Improvements and Developments in Gene Regulation and Single-Cell Gene Expression Data Analysis

by

Christopher Lee

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2020

Doctoral Committee:

Associate Professor Maureen A. Sartor, Co-Chair
Associate Professor Xiang Zhou, Co-Chair
Assistant Professor Alan P. Boyle
Associate Professor Hyun Min Kang
Research Professor Laura J. Scott

Christopher Lee

leetaiyi@umich.edu

ORCID iD: 0000-0002-8621-256X

DEDICATION

Shoutouts to Super Smash Brothers Melee for the Nintendo Gamecube

ACKNOWLEDGEMENTS

Thanks to my M.S. cohort, my friends in the Statistical Genetics group, my Sartorlabmates, and the MIDAS single-cell working group.

Additional gratitude for the Michigan Melee scene and the friends I met on Monster Super League

PREFACE

This is the culmination of years of training ~ Mòrag Ladair

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abstract

Recent advancements in high-throughput sequencing technologies have led to a vast amount of data assessing genome-wide regulation and single cell transcriptomics, which aid in the study of tissue complexities and heterogeneity at the molecular level among cells. These data provide opportunities for new insights into intracellular genetic mechanisms and require the development of new methods.

We first improve and extend methods for pathway analysis for large sets of genomic regions, such as those derived from genome-wide transcription factor (ChIP-seq) or open chromatin (e.g. ATAC-seq) experiments. Starting with ChIP-Enrich, a previously-developed gene set enrichment method, we improve the runtime by using an approximation to reduce redundant calculations. We then created Poly-Enrich, an extension of ChIP-Enrich to allow for count and weighted count outcomes per gene (number of genomic regions) as opposed to only binary outcomes. Comparing the results from ChIP to Poly-Enrich, we discover patterns in gene regulation based on transcription factor and gene functionality, to be used in predicting the more powerful method. Furthermore, we introduce a hybrid test to combine the two methods when the prediction of the more powerful method is ambiguous. Using Poly-Enrich, we evaluated several ways of defining enhancer locations and assigning them to target genes, and found several ways to improve the accuracy of distal regulation analyses.

Second, to complement gene set enrichment tests that do not account for relative peak locations, we developed a new method, proxReg, that tests whether experimentally-identified genomic regions that are associated with a specific

biological function tend to be farther or closer to regulatory regions, either gene transcription start sites or enhancer regions. Using proxReg alone, we find that transcription factors such as NRSF can bind closer to either type of regulatory region depending on the regulated biological processes. Complementing Poly-Enrich, proxReg provides additional insight into how a pathway is regulated, as well as providing additional evidence for its significance.

For the third project, we extend an existing method, PQLseq, which tests differential expression in bulk RNA-seq data with a mixed effects generalized linear model estimated by maximizing a penalized quasilikelihood, to the single cell RNA-seq (scRNA-seq) setting. In particular, we extend PQLseq to allow it to be used in hierarchical modeling structures with scRNA-seq data. A hierarchical data structure is very common in single cell studies. However, many existing scRNA-seq differential expression methods do not take this into account. Differential expression methods that can model hierarchical data structures are often computationally slow due to the handling of very large matrices in the data. Here, we take advantage of several properties of the hierarchical structure of single cell RNA-seq data and develop a new algorithm that allows for faster differential expression analysis. Our method specifically takes advantage of the natural low-rank structure in the data, saving several magnitudes of calculation time. With extensive simulations, we show that our method provides well controlled type I error and can provide higher power without astronomical computation times.

Chapter 1 Introduction

More than the genotype

When thinking about genetics, a common misinterpretation is to only consider the genotype, and downstream effects are only based on genic and intergenic sequences. However, evolution has devised a variety of mechanisms leading to the diverse cellular phenotypes among cells with identical genotypes. As DNA must first be transcribed to RNA, and then translated to proteins, there remain several opportunities to alter the expression of a gene. For instance, some proteins bind to the DNA to promote the transcription process, while others bind to disrupt it. Such proteins, called transcription factors (TFs), specifically regulate the transcription step of expression (Latchman 1997);(Karin 1990). As another example, histone binding can physically alter the structure of chromosomes, making it easier or harder for proteins to bind at their targets (Kouzarides and Bannister 2011);(Dong and Weng 2013). Several high-throughput assays have been developed to be able to measure various stages of protein formation and their function, requiring statistical methods to properly analyze such data, while accounting for their imperfections.

RNA-seq to measure gene expression

The most popular way to measure the activity of a gene is to directly measure the quantity of RNA transcripts formed. RNA-seq has been performed for over a decade on bulk tissues that contain many cells from potentially heterogeneous cell types. However, recent technological wet lab advances have enabled RNA-seq to be performed on single cells instead, called single-cell RNA-seq (scRNA-seq) (Eberwine et al. 2014). A bulk tissue containing several cells tends to be an amalgamation of

several cell types, cells in different stages of differentiation or cell cycle, or cells from external contamination. Separating each cell individually with scRNA-seq provides opportunities for questions that were formerly impossible, such as differential expression between cell types or limiting an analysis to one type of cell. It also provides a much larger effective sample size as one tissue sample may contain thousands of relevant cells, allowing us to characterize the cell heterogeneity within the tissue (Vieth et al. 2019). However, scRNA-seq assays are imperfect, causing challenges in the data analysis compared to that of bulk tissue. For example, some types of scRNA-seq technology suffer from lost counts due to dropout events (Kharchenko, Silberstein, and Scadden 2014). Occasionally, there are strong outliers due to accidental reading of multiple cells (McGinnis, Murrow, and Gartner 2019);(Bais and Kostka 2019) or sudden bursts of transcriptional activity (Larsson et al. 2019), which along with other factors, generally results in a much noisier data distribution due to the low sequencing depth and low capture efficiency.

With single cell expression data, one may be interested in differential expression (DE) analysis, which tests for a difference in expression level, e.g. between a case and control group or between cells designated as different cell types. Due to the aforementioned issues, scRNA-seq DE methods have to account for several more complexities in the data than bulk RNA-seq DE methods. Consequently, naively using bulk RNA-seq methods will likely cause higher error rates (Soneson and Robinson 2018). Additionally, a much larger sample size also restricts more complex analysis as the time required to perform such analyses starts to increase to impractical amounts of time. For example, for a study where there are individuals in groups, and each individual has single cells, i.e. a hierarchical data structure (Figure 1.1), it would be ideal to use a mixed effects model to control for individual effects. Unfortunately, the fitting algorithms for mixed effects models do not scale well with larger data (>10,000 cells), so few methods attempt modeling the hierarchical data structure in scRNA-seq data. Therefore, one of the challenges for differential

expression analysis in scRNA-seq is to find the optimal balance between power and computational cost.

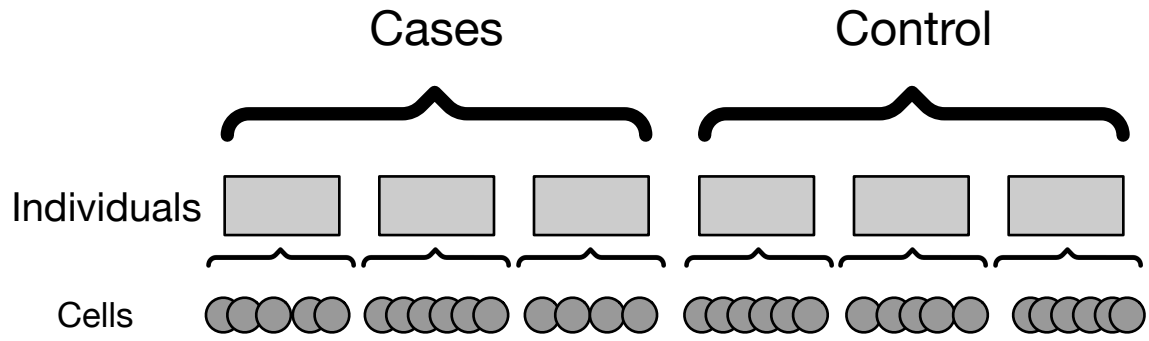


Figure 1.1 An example of a hierarchical data structure in scRNA-seq, where there are individuals within each case and control group.

Each individual has a sample of cells, generally thousands of cells per individual. We would assume the properties of cells in each individual are more correlated with each other than with cells from other individuals.

Statistical methods for differential gene expression, and complexities with scRNAseq

Before the discovery of scRNA-seq, several differential expression analysis models for bulk RNA-seq counts were already well developed. Similar to any other count-based model, many are based on a generalized linear model with an overdispersion parameter (i.e. negative binomial). Due to aforementioned issues in scRNA-seq that are not present in bulk RNA-seq, some methods use different strategies to account for the complexities in the data. One such adjustment is to include a zero-inflation parameter to account for the abundance of zeros from dropout events, as implemented in Monocle (Trapnell et al. 2014) or using observation weights from zingeR (Van den Berge et al. 2018) accompanying EdgeR (Robinson, McCarthy, and Smyth 2010) or DESeq (Love, Huber, and Anders 2014). Additionally, one may extend this and model the dropout parameter in a multi-stage model as in SCDE (Kharchenko, Silberstein, and Scadden 2014) or MAST (Finak et al. 2015).

There are several philosophies on how to approach scRNA-seq data analysis: some argue that modeling zero-inflation is unnecessary (Svensson 2020), some use a linear approach instead of a count-based approach (Finak et al. 2015), some test for difference in distribution (Korthauer et al. 2016), while still others may be content with a simple t-test or Wilcoxon test (Soneson and Robinson 2018). While no method may be inherently superior to every other in all situations, it is important to distinguish the scenarios where one method is more powerful than another, and methods that can robustly cover the most scenarios.

Transcription factor binding sites and other genomic regions

A transcription factor (TF) may bind to the DNA in tens of thousands of locations, often near a gene's transcription start site (TSS). While the locations of genes and their TSS's are well known, where and why the TFs bind are less known. Using chromatin immunoprecipitation and sequencing (ChIP-seq), we are able to identify the locations where a TF binds during any biological context of interest in the form of read pile-ups (i.e. peaks). ChIP-seq, in summary, finds DNA fragments that are bound by the TF of interest across the genome in many cells in a tissue. While the TFs are bound to the DNA, the fragments are isolated by cutting the DNA and using an antibody specific to the TF to obtain only the TF-DNA fragment interactions. After removing the TFs bound to the fragments (usually in the range of 200-400 bp), one or both ends of these fragments are then sequenced and mapped to the relevant reference genome so that the location and number of read fragments generate the peaks and their properties. For instance, a sequence of base pairs that is a very common TF binding site would result in several DNA fragments mapped to that part of the genome and would be interpreted as a peak of high intensity. We can then use these properties, such as the number, location, width, and/or intensity of these peaks to discover associations between them and the types of genes that are regulated. Such elements that can be directly mapped to the genome, i.e. defining exactly where on a chromosome with start and end coordinates, are called genomic regions. Another particularly useful type of genomic region is a motif, which is a

short, recurring pattern in the genome that indicates where a TF binds, and can be detected near the center of ChIP-seq peaks (D'haeseleer 2006) (Figure 1.2). Other examples of sets of genomic regions of interest are repetitive elements (de Koning et al. 2011) or CpG sites (Jabbari and Bernardi 2004), and all of them can be used to help explain the functionality of certain genes.

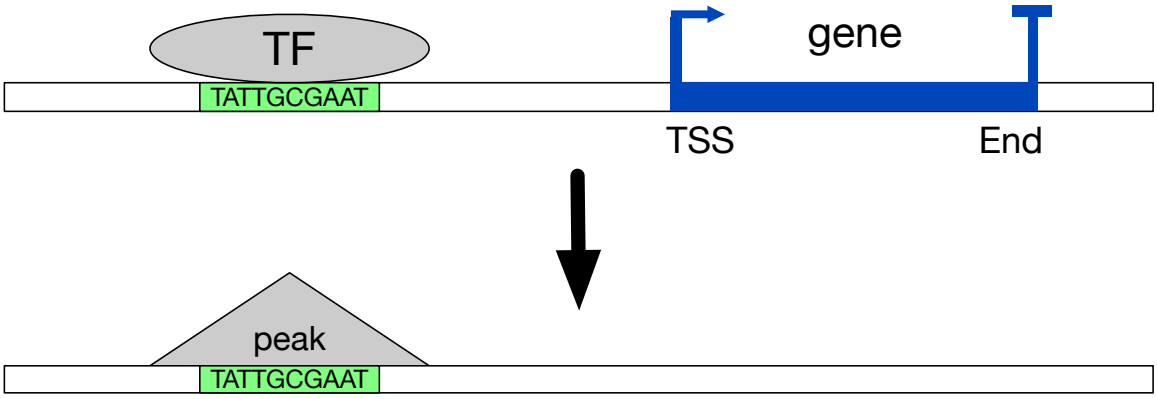


Figure 1.2 An example of a TF binding to its motif near and upstream from a gene's transcription start site (TSS), called the promoter region.

This TF binding is then interpreted by ChIP-seq as a peak (generally 200-250 base pairs wide compared to a motif being around 10 base pairs), with properties such as height that usually corresponds to binding strength, and width that corresponds to binding coverage.

Regulation near and far from genes

We refer to the region immediately around a gene's TSS, around 10-1000 bases upstream (Sharan Lecture 11, January 4, 2007 2007 #350), as the promoter region. TFs that bind to these regions are assumed to directly regulate that gene as its target gene, as promoter binding affects the ability of RNA Polymerase to bind and transcribe the gene (deHaseth, Zupancic, and Record 1998). However, TFs can bind in locations far from their target gene, termed distal binding, which instead can affect the physical mechanisms of a DNA strand looping on itself (Stadhouders et al. 2012). Distal binding TFs can increase (enhance) or decrease (silence) gene transcription, and the locations they bind to are referred to as distal regulatory elements (DREs), specifically either enhancers or silencers (Figure 1.3). It is possible

for there to be several genes in between distal regulating TFs and their target genes, and it is not obvious which gene an arbitrary distal TF binding event is regulating.

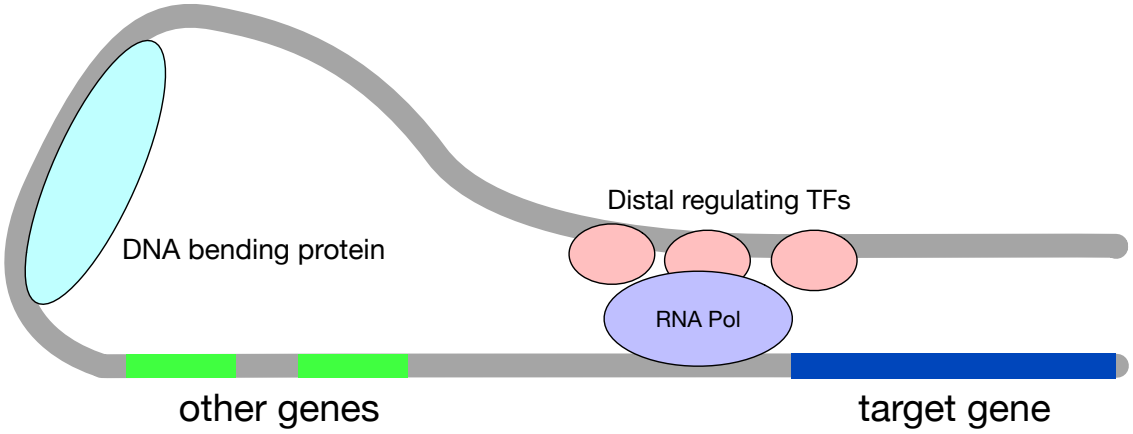


Figure 1.3 An example scenario of distal binding proteins affecting gene transcription via directly affecting RNA Polymerase's activity.

Some TFs can directly bend the structure of DNA (blue), which allows the distal binding TFs (pink) to increase or decrease the activity of RNA Pol II. Distal regions where TF binding increases activity are called enhancers and those that decrease activity are called silencers. There also tend to be several other genes in between the distal TFs and their target gene, so it is difficult to correctly identify a distal TF's target gene.

Analysis by considering a set of genes

Genes and their coded proteins rarely work alone, but instead in a biological pathway containing several to many hundreds of genes. Thus, we can analyze pathways in the form of gene sets using gene set enrichment (GSE) tests, where we say a gene set is *enriched* if the outcome of interest has a significantly larger signal in the genes in the gene set, and we say a gene set is *depleted* for the opposite direction. Gene sets can be user defined, but there are several existing databases such as the Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto 2000) Pathways and Gene Ontology (Harris et al. 2004), all with studies to support the gene set assignments.

Most GSE methods were developed for gene expression data, either microarray or RNA-seq (Slonim and Yanai 2009, Chu and Corey 2012). If one has a gene-level dataset where the outcome of interest is binary, one could then create a 2x2

contingency table between the outcome and gene set membership. In this case, one could use a Fisher's Exact Test or a modification, such as in the commonly-used GSE testing software, DAVID (Huang et al. 2007). With RNA-seq data, one can measure the level of a gene's differential expression, as in RNA-Enrich (Lee, Patil, and Sartor 2016). There are also several other types of outcomes for genomic data, so many more methods have been developed to cater to different scenarios.

Statistical methods for gene set enrichment testing of large sets of genomic regions

For ChIP-seq data, one can measure the proportion of genes in a pathway that have at least one peak as is done in a previous method, ChIP-Enrich (Welch et al. 2014), the number of peaks assigned to genes in a pathway as in the method GREAT (McLean et al. 2010), or the coverage of the gene loci in a pathway as in Broad-Enrich (Cavalcante et al. 2014). Depending on interest, peak-to-gene assignment can be chosen to focus on specific parts of the genome, such as only the promoter regions or exons, or all peaks can be assigned to their nearest gene TSS. It is also possible to perform GSE tests on other types of genomic regions, including ATAC-seq peaks, repetitive element families, differential DNA methylation sites, or single nucleotide polymorphisms from genome-wide association studies. Specifically, the last can be performed by SSEA (Weng et al. 2011), i-GSEA4GWAS (Zhang et al. 2010), MAGENTA (Segrè et al. 2010), or GSA-SNP2 (Yoon et al. 2018).

Gene set enrichment tests for sets of genomic regions are most often *competitive* tests, meaning each gene set effectively 'competes' with other gene sets for significant enrichment. The null hypothesis is that genes in the gene set have at most as much signal from genomic regions as those not in the gene set (Goeman and Bühlmann 2007). In contrast, *self-contained* tests only use information from the genes in the gene set, testing the null hypothesis that the signal from the genomic regions in the gene set is not significantly more than zero.

Chapter Overview

In this dissertation, we develop several methods to analyze data that measure gene regulation and expression to be able to better understand the mechanisms beyond the genotype. In Chapter 1, I introduce Poly-Enrich, which is a competitive GSE test that discovers the effects of the quantity of genomic regions in a pathway. In Chapter 2, I introduce proxReg, which facilitates interpretation of Poly-Enrich or any other GSE test results for genomics regions, by adding insight into the relative locations of the genomic regions, that is, whether they are more proximal to promoter or enhancer regions than expected by chance. Finally, Chapter 3 introduces a low-rank PQLseq extension to perform differential expression analysis on a common, but complex data structure in scRNA-seq experiments.

CHAPTER 2 Poly-Enrich: Count-Based Methods for Gene Set Enrichment Testing with Genomic Regions

A paper covering most of material in this chapter is in review at NAR: Genomics and Bioinformatics, with myself as first author.

A paper covering the enhancers portion of this chapter is in preparation, with myself as second author.

Introduction

Regulatory genomics experiments help us understand how cells use more than their genetic sequence to carry out a vast repertoire of cellular programs. Common regulatory genomics methods include chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-Seq) and ATAC-seq, which identify transcription factor (TF) binding sites and open chromatin regions, respectively, across the genome. Other types of data, such as DNA Methylation assays, copy number alterations, repetitive element families and groups of SNPs, also lead to large sets of genomic regions that potentially play a specific role in regulatory genomics, with each type having notably different properties in terms of the number, size, and location of genomic regions.

Proteins that bind near a gene can regulate it in ways such as improving structural properties or physically blocking other proteins, often positively or negatively regulating the gene's expression, respectively. Additionally, some proteins bind DNA several times in a clustered region (Gotea et al. 2010), or in distant enhancer regions that interact with the same or distinct proteins bound in promoter regions

(Pennacchio et al. 2013). Binding sites also differ in strength; a protein may bind in only a portion of cells in a sample at the time of immunoprecipitation, either due to weak binding or due to varying chromatin accessibility among the cell types in the sample. These binding sites along the genome are interpreted as peaks of varying strengths, depending on the signal-to-noise ratio or significance level of the peak. In general, interpreting each peak's target gene(s) and effects remains an active area of research, which over time may improve results on downstream tests such as gene set enrichment.

With so many different tests, one may wonder which test is optimal for their data, but there is no single recommendation across data types. Different tests are needed for different types of genomic regions as properties such as peak widths, number of peaks, and location relative to genes all make a difference. Thus GSE testing for genomic regions should not be a one-size-fits-all test; some methods work better than others in specific scenarios. For example, Cavalcante et al. showed that Broad-Enrich is more powerful than ChIP-Enrich for broad regions, but lacks power for narrow regions (Cavalcante et al. 2014). As another example, GREAT does not account for variability among genes, so it is best used in situations where the probability of a peak is constant across genomic space (e.g. per kb), as opposed to clustered near transcription start sites or displaying variability among gene loci.

Our previous method, ChIP-Enrich, uses a binary score to classify a gene as having at least one peak. We saw that ChIP-Enrich tends to underperform when nearly all genes have at least one associated genomic region; in this case, ChIP-Enrich will not yield meaningful results. We hypothesized that a count-based method that captures the frequency of binding would perform better in those situations. In this paper, we introduce such a method, Poly-Enrich to expand our available methods to be suitable for any set of narrow genomic regions including those that tend to saturate genes. The flexible structure of the Poly-Enrich test also allows additional capabilities, such as accounting for the strength of each ChIP-seq peak. Whereas ChIP-Enrich has the hypothesis that a single binding site is sufficient for regulation,

Poly-Enrich allows for regulation that is incremental, i.e. more genomic regions correspond to stronger or more likely regulation. To identify under which situations one is more appropriate than the other, we performed a comparison of Poly-Enrich and ChIP-Enrich using a set of 90 transcription factor (TF) ChIP-seq datasets from the Encyclopedia of DNA Elements (ENCODE) (Rosenbloom et al. 2010). We also introduce a hybrid test that combines information from both ChIP-Enrich and Poly-Enrich.

To illustrate the usage of Poly-Enrich, we apply it to sets of repetitive elements in the human Alu and LINE1 families, revealing for the first time a comprehensive view of the processes and functions enriched or depleted with these repetitive elements in the human genome. Finally, we describe several updates to our ChIP-Enrich website and *chipenrich* Bioconductor package, including additional methods for assigning genomic regions to target genes, new gene set databases, and more supported species.

Methods

Datasets

All ChIP-Seq data were obtained from Encyclopedia of DNA Elements (ENCODE) at University of Califonia, Santa Cruz (Rosenbloom et al. 2010). We used a total of 90 experiments over the three Tier 1 cell lines (Gm12878, H1-hESC, and K562), and all 35 transcription factors that had available ChIP-seq data for at least two of the three Tier 1 cell lines (Supplementary Table 1).

The gene sets used were from Gene Ontology: Biological Processes (GOBP) ver. 3.4.2 (Ashburner et al. 2000). We filtered out gene sets with less than 15 genes or more than 2000 genes as gene sets with too few genes generally have insufficient power and may not satisfy the assumptions of the statistical model, and gene sets with too

many genes are too vague to be biologically informative. In total, there were 5015 gene sets used.

Assigning regions to genes

The UCSC knownGene database for hg19 was used to define the transcription start sites across the genome (Hsu et al. 2006). Each gene locus definition (e.g. nearest TSS, <5kb, etc) was generated as a table containing the columns: chromosome, Start, End, gene ID.

Poly-Enrich model: a generalized linear model with a negative binomial family

We model the number of genomic regions assigned to each gene using a generalized linear model (GLM) with a negative binomial (NB) family. The model is:

$$\ln(\mu_i) = \beta_0 + \beta_1 GS_i + f(\ln(LL_i * m + 1)$$

where for each gene i, GS is an indicator for whether the gene is in the gene set of interest or not (=1 if in the gene set; 0 otherwise), μ is the mean of the negative binomial distribution for the number of genomic regions assigned to each gene, and the overdispersion parameter θ is simultaneously estimated so that $Var(Y|GS) = \mu + \theta\mu^2$, where Y is the number of genomic regions for the gene. We plotted mean vs variance of the counts per gene for our data sets and confirmed that the relationship is close to a quadratic (data not shown). The function f is a cubic smoothing spline that adjusts for the gene's locus length and optionally adjusts for *m,* the mappability of the gene's locus. Details about how we adjust for mappability can be found in the ChIP-Enrich manuscript (Welch et al. 2014). We use the *gam* function in the *mgcv* R package to fit the model, which uses a penalized likelihood maximization, and the smoothing spline penalty is a squared second derivative penalty (Wood, Goude, and Shaw 2015). Use of a cubic smoothing spline to adjust for the genes' locus lengths was first introduced in ChIP-Enrich, and has been shown to be a powerful, flexible way to model this relationship (Welch et al. 2014).

A likelihood ratio test (LRT) on the coefficient for the gene set is used to test for enrichment (or depletion) of each gene set: the test statistic is defined as $L = -2(l_0 - l_1)$, where L follows a χ_1^2 distribution under the null hypothesis that there is no association between gene set membership and number of genomic regions (i.e. $\beta_0 = 0$), and l_0, l_1 are the maximum log likelihoods under the null and alternative hypotheses, respectively. We use the LRT instead of the Wald test, because the LRT was shown to perform significantly better than the Wald test with generalized linear models using a negative binomial family (Robinson and Smyth 2008). We then look at the sign and significance of β_1 to test for enrichment, where a positive β_1 indicates enrichment, and a negative value indicates depletion (fewer regions than expected at random). For each gene set of interest, we estimate a different set of model parameters, and correct for multiple testing afterwards.

Poly-Enrich with weighting based on genomic region scores

In certain cases, each genomic region in a dataset may be associated with a numeric score. For example, ChIP-seq peak finding results often include a value denoting the strength of a peak, (e.g. signalValue in ENCODE dataset results or -10*$\log_{10}$(p-value) in MACS2 results). Poly-Enrich weights based on these scores by giving each genomic region a weight proportional to its signal value (or other score) and normalizing such that the mean of all weights is equal to 1. For every genomic region assigned to a gene, we sum all the weights and substitute the weighted sum in place of the original count. The same model can still be used on non-whole number data as the calculations are equivalent while using the Gamma function instead of a factorial.

Comparing p-values between methods

To compare p-values between methods, we use a scatterplot, plotting a signed -$\log_{10}$ p-value per gene set. If a gene set is enriched, the sign is positive, and if the

gene is depleted, the sign is negative. This allows us to detect if there are any cases where two methods may contradict each other's conclusions.

Spline approximation for Poly-Enrich and ChIP-Enrich

With a library of over 20,000 genes and most gene sets being less than 1000 genes, the cubic smoothing spline estimate changes very little between gene sets. Thus, we have confirmed we can reasonably assume that the spline is approximately equal for any gene set of interest, including the spline with no gene set (Supplementary Figure 2.1A, B).

We first run the same model except without the gene set (*GS*) term: $\ln(\mu_i) = \beta_0 + f(LL_i)$. We then extract the fitted spline using the *predict* function with *type="terms"* from the *mgcv* R package to obtain a spline-adjusted locus length for each gene. This new value is then input as a covariate in the model for every gene set, which allows us to fit a spline only once instead of once for each gene set. This saves a significant amount of time when testing a large number of gene sets (approximately 75% time saved when testing 4000 gene sets). Compared to the original model, we find that the -log p-values from the spline approximation model are nearly identical (Supplementary Figure 2.1C, D).

We also considered using the spline estimate as an offset instead of a covariate. We found that the coefficient for the spline term is close to 1 in almost all cases and results were practically identical (data not shown).

Score test for fast estimates

Compared to the likelihood ratio test, the score test requires the least amount of computation time in exchange for lacking power in negative binomial families (Robinson ref). However, the score test is still a good for situations when one needs a large amount of preliminary results. We use the *glm.scoretest* function from the

statmod (Giner and Smyth 2016) package to compute the score test for both ChIP-Enrich and Poly-Enrich. The score test runs over 30 times faster and is reasonably concordant with the LRT for enriched GO terms, but less similar for depleted GO terms. (Supplementary Figure 2.2)

Testing Type I error

The null hypothesis of Poly-Enrich is that there is no true biological enrichment. To test the Type-I error, we randomly permuted the genes to simulate scenarios where there is no association between genes and the number of peaks. However, to ensure that the results are not biased by gene locus length or gene location, we performed two additional permutations: one permutes genes within bins of similar locus length, while the other permutes within bins of chromosomal locations. In both cases, the genes are sorted by the variable of interest (locus length or location), and then assigned to consecutive bins of 100 genes each. These randomization tests are identical to those used in the Broad-Enrich manuscript (Cavalcante et al. 2014).

For each of the 90 TF peak datasets chosen, after assigning the peaks to genes, we permuted the gene IDs using the randomization of interest, and then performed enrichment tests against GO biological processes. We ran a total of 10 trials and took the median p-value per gene set as the randomization p-value. Then, the proportion of p-values less than a defined confidence level was determined per experiment to calculate the overall Type I error. We then plotted all 90 overall Type I errors for each experiment in a box plot to convey overall Type I error.

Testing power

To test statistical power, we chose three TF peak data sets of varying size (4194, 11129, and 40052 peaks) and two gene sets of varying size (42 and 471 genes) as our base scenarios. To illustrate how Poly-Enrich can detect enrichment for data sets with very large numbers of peaks (beyond what ChIP-Enrich can handle), we included two larger data sets: an ATAC-seq data set with 99,478 genomic regions, and an Alu repetitive elements data set with 1,094,736. After assigning the genomic

regions to genes, we randomized the genes in bins of locus length to remove all true gene set enrichment signal while keeping locus length association, and then randomly added peaks into the gene set to simulate enrichment. We chose three scenarios of enrichment, each with varying levels (*x%*=5, 10, 20, or 30) of enrichment:

1. *CEbias*: Enriched to closely satisfy the assumptions of the binary (ChIP-Enrich) model. We added peaks to *x*% of the remaining genes in the gene set without a peak. This increases the proportion of genes with a peak, without causing a large increase in the mean number of peaks per gene.

2. *PEbias*: Enriched to closely satisfy the assumption of the count-based (Poly-Enrich) model. We added a number of peaks, equal to *x*% of the number of peaks in the gene set, to one twelfth of the genes in the gene set. This increases the mean number of peaks per gene, with little effect on the proportion of genes with a peak.

3. *Balanced*: We added a number of peaks, equal to *x*% of the number of peaks in the gene set, into the gene set weighted by gene locus length. This increases both the proportion of genes with a peak and the mean number of peaks per gene by a similar degree.

Defining the true positive transcription factor-GO term pairs

For each transcription factor, we identified the gene that codes for it from R package *GO.db* (Carlson 2019), and then identified every GO biological process that gene is assigned to. Pol2 is excluded as its functions are too widespread, resulting in 25 transcription factors, each of which is assigned to at minimum 50 GO BP terms. This set of GO terms, along with its parents and grandparents, is what we use as the true positive set. We also define a true negative set of GO BP terms as every other GO term, except: ancestors, siblings, and offspring of the true positive set, and terms with ≥ 2000 or ≤ 10 genes. Using the true positive and negative sets, we calculated

empirical false positive rates (FPRs) for Poly-Enrich, GREAT, and ChIP-Enrich. This estimated FPR serves as an upper bound for the true FPR as it is not a perfect gold standard (i.e. some negative GO BP terms may actually be novel true findings, since some functions of a TF may be unknown).

Hybrid test

The hybrid method introduced by Zhang et al (16), which we employ, was shown to be especially beneficial when there is no one optimal test in all cases. Given *n* tests that test for the same hypothesis, the same Type I error rate, and converted to p-values $p_1, \dots, p_n$, the Hybrid p-value is computed as:
$p_{hybrid} = n \times \min(p_1, \dots, p_n)$, which is the same as a Bonferroni adjustment for multiple p-values. This hybrid test will have at most the same Type I error rate as the *n* tests, and if at least one test is consistent (power converges to 1 as sample size reaches infinity), the hybrid test will also be consistent. Proofs and simulations of the test in general were done by Zhang et. al (Zhang et al. 2016). Here, we've implemented the hybrid test for users to use two methods ($n = 2$): ChIP-Enrich and Poly-Enrich. Users may also choose any two results files and run a hybrid test based on those.

Clustering and heatmaps

For every GO term, we calculated the difference in $-\log_{10}$ p-value for each of the 90 experiments between ChIP-Enrich and Poly-Enrich, with positive values indicating a more significant result for Poly-Enrich. We then focused on GO terms where > 10% of the experiments had an absolute $\log_{10}$ p-value difference greater than 2.
Clustering was performed using uncentered correlation as the similarity metric and average linkage as the clustering method. Using Java TreeView, we extracted specific groups of GO terms that contain certain strings such as "cell cycle" or "positive regul."

Repetitive elements

Data was obtained from the UCSC Table Browser with RepeatMasker 3.0 on the hg19 genome. We chose the two most abundant families in the dataset: Alu and L1, as well as four methods of peak-to-gene assignments: Intron, Nearest TSS, >5kb, and <5kb. Poly-Enrich was then used to perform gene set enrichment. Before clustering for the heatmap, we filtered out GO terms where there were 2 or fewer significant FDR values among the 8 categories. The clustering method was the same as mentioned in the previous section.

Website and Bioconductor updates

The Chip-Enrich website (http://chip-enrich.med.umich.edu) was updated from the *chipenrich* package version 1.7.2 to version 2.5.0. (from https://github.com/sartorlab/chipenrich, on Aug 8th, 2018). We have added the following reference genomes: human (hg38), rat (rn5, rn6), *Drosophilla melanogaster* (dm6) and zebrafish (danRer10). We also added the following databases from MSigDB (Version 6.0): Hallmark, Immunologic, MicroRNA, Transcription Factors, and Oncogenic (Liberzon et al. 2011, Liberzon et al. 2015), and sets of genes that are known to be affected by particular environmental toxins from the Comparative Toxicogenomics Database (CTD) (Davis et al. 2017). We also provide direction in the vignette for how to use gene sets from other R packages, such as EGSEAdata (Alhamdoosh et al. 2017).

In addition to the previous locus definitions ('nearest TSS', 'nearest gene', '≤1 kb from TSS' and '≤5 kb from TSS'), we also now support gene locus definitions for regions <10 kb from a TSS and gene distal regions (>10kb upstream of a TSS).

Results

Motivation for development of Poly-Enrich

The motivation for our new methods comes from situations observed with real sets of genomic regions, often with ChIP-seq peak datasets, but also from other sources, such as families of repetitive elements or large sets of DNA polymorphisms such as those different between closely related species or sub-species. Although our original method, ChIP-Enrich, performs extremely well for most transcription factor (TF) ChIP-seq datasets (Figure 2.1A), because it uses a simple binary score for each gene, there are some scenarios where this simplification has a significant loss of information. For example, ChIP-Enrich models a gene with many peaks the same as a gene with only one peak, even though gene regulation may be affected by additional peaks (Figure 2.1B). Alternatively, if nearly every gene is assigned at least one peak, ChIP-Enrich would be unable to distinguish among them and thus unable to detect any gene set enrichment (Figure 2.1C).

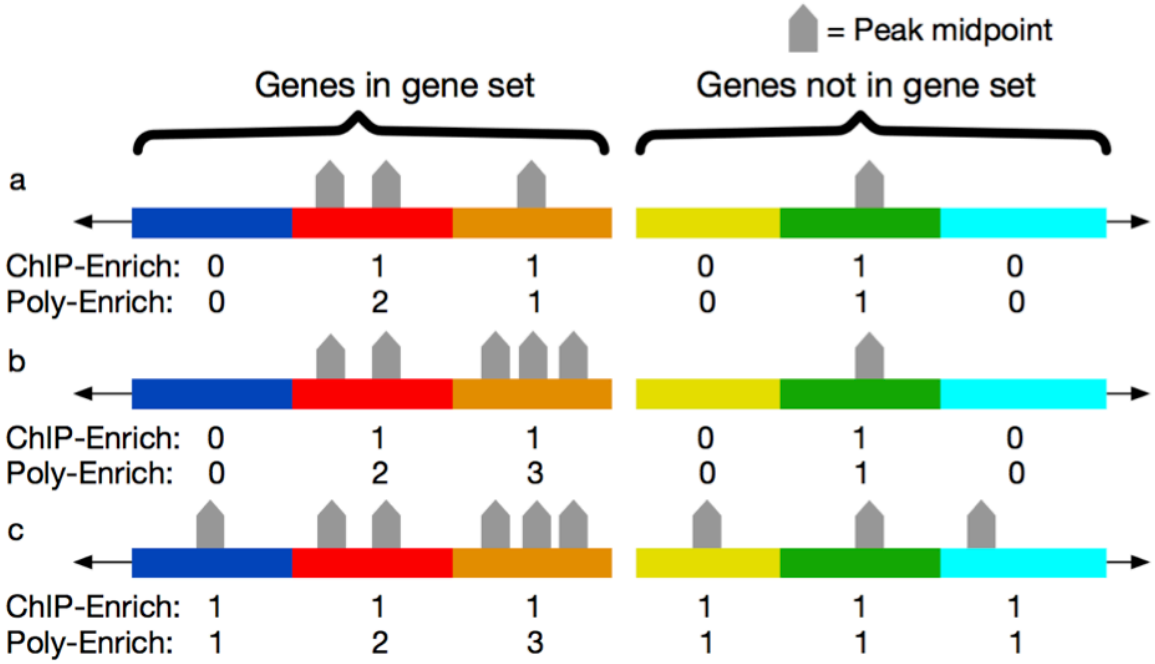


Figure 2.1 Three scenarios of ChIP-seq peak distributions illustrating how ChIP-Enrich and Poly-Enrich perform.

Each color represents a different gene locus; the left three are in a gene set and the right three are not. (A) Peaks are relatively evenly distributed, with a small number

across a subset of genes. Given this situation, ChIP-Enrich evaluates 2/3 vs 1/3 while Poly-Enrich evaluates [0,2,1] vs [0,1,0]; both methods perform well. (B) Some genes contain significantly more peaks than others, such that information is to be gained from the number per gene. ChIP-Enrich evaluates 2/3 vs 1/3, Poly-Enrich evaluates [0,2,3] vs [0,1,0]; ChIP-Enrich performs adequately, but Poly-Enrich is optimal. (C) Nearly all genes have at least one peak, with some having significantly more than others. ChIP-Enrich evaluates 3/3 vs 3/3, Poly-Enrich evaluates [1,2,3] vs [1,1,1]; ChIP-Enrich would not detect any enrichment, while Poly-Enrich can still detect gene sets enriched with more peaks.

Although the alternative current approach, GREAT, is also a count-based gene set enrichment method, Poly-Enrich differs significantly from it in two respects. Firstly, whereas GREAT counts the number of peaks in an entire gene set, Poly-Enrich counts them per gene. By separating counts per gene, we are able to adjust for each gene's locus length and the variability in peak count across genes, which we previously showed was an important adjustment to control for Type I error (Welch et al. 2014). Secondly, the binomial model used by GREAT assumes that the background probability of a peak is constant across the genome. Poly-Enrich uses a more flexible, empirical approach to this that provides for a range of different assumptions about peak distribution. As previously shown, the consequences are that GREAT does not provide accurate significance estimates (the resulting p-values are more significant than they ought to be), and it tends to rank gene sets with shorter genes more highly than those with longer genes (Welch et al. 2014). We therefore developed Poly-Enrich as a count-based competitive method that addresses all of the above mentioned shortcomings of ChIP-Enrich and GREAT.

ChIP-Enrich, GREAT, and Poly-Enrich all use a region's midpoint to define its location. These genomic regions can then be assigned to genes in different ways, so that regulation from different types of regions (e.g., promoters, introns, or regions distal to TSSs) can be studied. We define a gene's locus definition as the region on the genome such that peaks in that region are assigned to the gene. These loci are defined using properties of the gene, such as within 5kb of a gene's transcription start site (TSS), or simply by assigning each region to the nearest TSS (Figure2.2). In

the new version of our GSE website and *chipenrich* Bioconductor package, we offer several additional choices, including exons, introns, and distal regions only (>10kb upstream from a TSS). Users can also upload their own custom locus definition, such as open chromatin regions for a specific cell type, or known enhancers and their target genes.

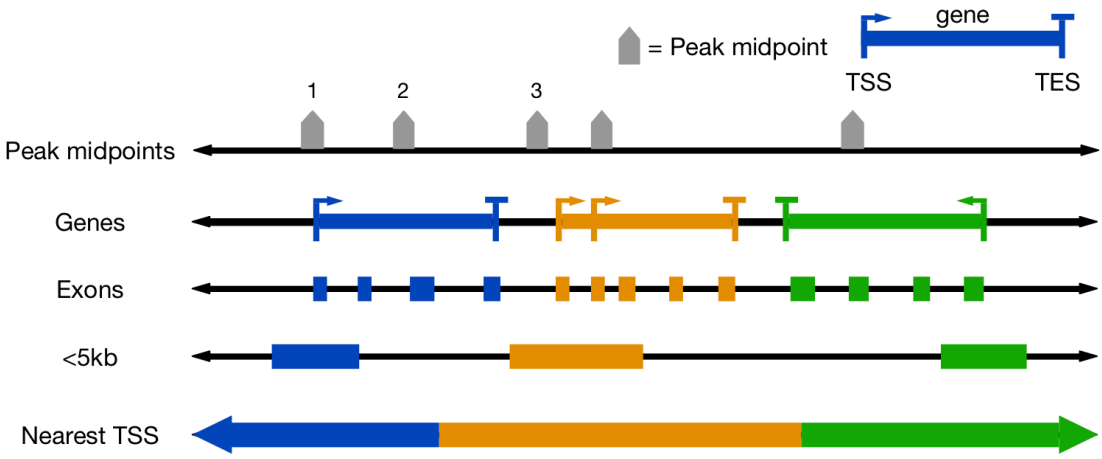


Figure 2.2 Overview of peak-to-gene assignments given gene locus definitions.

Given the gene locations and definitions for a genome, several different methods for assigning genomic regions to genes can be defined, referred to as gene locus definitions. Examples shown are: *Exons* – only peaks in any exon of a gene are assigned to that gene; *<5kb* - peaks within 5 kb of a gene's TSS are assigned to that gene; and *Nearest TSS* - peaks are assigned to the gene with the closest TSS. A gene's locus length is defined by the number of base pairs that could be assigned to the gene. In this toy example, peak 1 would be assigned to the blue gene for all three example gene locus definitions, peak 2 would only be assigned to the blue gene for the *Nearest TSS* locus definition, and peak 3 would be assigned to the orange gene only for the *<5kb* and *Nearest TSS* locus definitions.

Testing Type 1 error and power

We tested the type I error rate of the count-based method under the null hypothesis of no enrichment signal. By permuting the genes in the peak-to-gene assignment pairs and breaking the peak-gene relationships, we mimicked three scenarios of no enrichment: *i)* the "*complete*" randomization was done by shuffling the gene IDs in the whole dataset; *ii)* the "*bylength*" randomization was performed to verify that our method adequately adjusts for locus length, by first grouping genes into bins of similar locus length to preserve the locus length relationship; *iii)* the "*bylocation*"

randomization was performed to verify that the method adequately adjusts for relationships among genes in close proximity to each other, by grouping genes by their physical location to preserve relationships along the chromosomes. (See *Methods* for more detail.) We ran the randomizations on our 90 selected ChIP-Seq datasets from ENCODE (see *Methods*), and the proportion of p-values < 0.05 and < 0.001 for each dataset were plotted (Supplementary Figure 2.3A,B). We see that the test is slightly inflated but are at an acceptable level for Type 1 error in all cases. That is, approximately 5% had p-values < 0.05 and approximately 0.1% had p-values < 0.001 as expected. We observed a slight inflation in the "bylocation" randomization, which upon examination, we found to be caused by certain large clusters of functionally-related genes that are located near each other, for instance a cluster of histone genes which affected the results for Gm12878 ETS and H1hesc TBP (Supplementary Table 2.2). We previously showed that GREAT has an inflated Type 1 error under the "*complete*" and "*bylength*" randomizations, also using ENCODE ChIP-seq data (Welch et al. 2014).

To characterize the statistical power of Poly-Enrich under different situations, we permutated data while simulating enrichment of a gene set, and compared the results with those from ChIP-Enrich. We used three datasets with a small, medium, and large number of peaks, and two GO terms with a small and large number of genes. Three types of enrichment were simulated: one that adds peaks mainly according to the regulatory assumptions of ChIP-Enrich (CEBias), one that adds peaks mainly according to the assumptions of Poly-Enrich (PEBias), and one that is balanced between the two assumptions. For each type of enrichment, we simulated four levels of enrichment: 0.05, 0.1, 0.2, and 0.3, which indicates the proportion of additional peaks added to the gene set. (See *Methods* for more detail.) Finally, we chose two different levels of significance: $\alpha = 0.05$ and 0.001, as our cutoffs.

As expected, higher simulated enrichment resulted in higher power, since adding more signal increases the ability of a test to detect significance. Also, larger gene sets have higher power due to an increased confidence in the estimated mean number of

peaks. Overall, we see that Poly-Enrich has more power in simulations that enrich a gene set by increasing the number of peaks per gene, while ChIP-Enrich has more power in simulations that enrich a gene set by adding peaks to genes without any previous peaks. Finally, the *Balanced* simulation results in the two methods having similar power in most cases (Supplementary Figure 2.4A,B).

With the two largest datasets, we tested power for the *Balanced* simulated gene sets to illustrate that Poly-Enrich is able to detect signal even when ChIP-Enrich fails. We see that ChIP-Enrich is can still perform reasonably well compared to Poly-Enrich with around 100k peaks, but starts being unable to detect any enrichment in the data set with over 1 million peaks where 81% of genes are assigned a peak in the small gene set, and actually loses power when more signal is added in the large gene set (Supplementary Figure 2.4C).

Validation with true positives

To complement our permutations and simulations, we compared Poly-Enrich, ChIP-Enrich, and GREAT's ability to find true positives while avoiding false positives with real ChIP-seq data. To do this, we first created a set of true positives comprised of GO term-TF pairs by using the GO term biological process (BP) assignments for the gene encoding the transcription factor (e.g. the gene encoding for JunD is assigned to the GO term, "cell death"). This makes the reasonable assumption that TFs tend to regulate genes in the same biological processes in which they are active. Out of the 25 TFs with at least 50 assigned GO terms, we found that GREAT had a larger empirical false positive rate (FPR) than both ChIP-Enrich or Poly-Enrich for 22 TFs (Supplementary Figure 2.5). Estimated FPRs were similar between Poly-Enrich and ChIP-Enrich, with 13 (52%) experiments being higher for ChIP-Enrich. The overall high FPR (compared to the expected 5%) can be attributed to the true positives being imperfect (see Methods).

Poly-Enrich with weighted genomic regions

The height and confidence of peaks in a ChIP-seq experiment can vary dramatically, thus we reasoned that incorporating this additional information would improve the ability to pinpoint the truly enriched pathways. Although the most apparent motivation for weighting genomic regions is to account for ChIP-seq peak strength, other situations exist where each peak or genomic region may be assigned a unique score (e.g. confidence or quality score). Due to the flexible nature of the Poly-Enrich model, we were able to easily add the option to weight regions by peak strength (using peak *signal value*; see *Methods* for details), and examined the extent to which adjusting for peak strength improves enrichment results using 90 ENCODE ChIP-seq datasets by comparing the $-\log_{10}$ p-values per gene set.

We noticed for 25% of the experiments, most enriched gene sets were more significant with weighting, thus as we hypothesized, binding events near genes in enriched GO terms were stronger than those near other genes Figure 2.3A, B). In another 20% of the experiments, the enrichment p-values were split between the two methods (Figure 2.3C). Interestingly, the distribution of $\log_{10}$ signal values for these experiments showed a bimodal pattern (Figure 2.3D). This suggests that some gene sets tend to have genes with significantly stronger binding peaks than others, and that both sets may be biologically interesting. For the remaining 55% of experiments tested, weighting made little difference on the results.

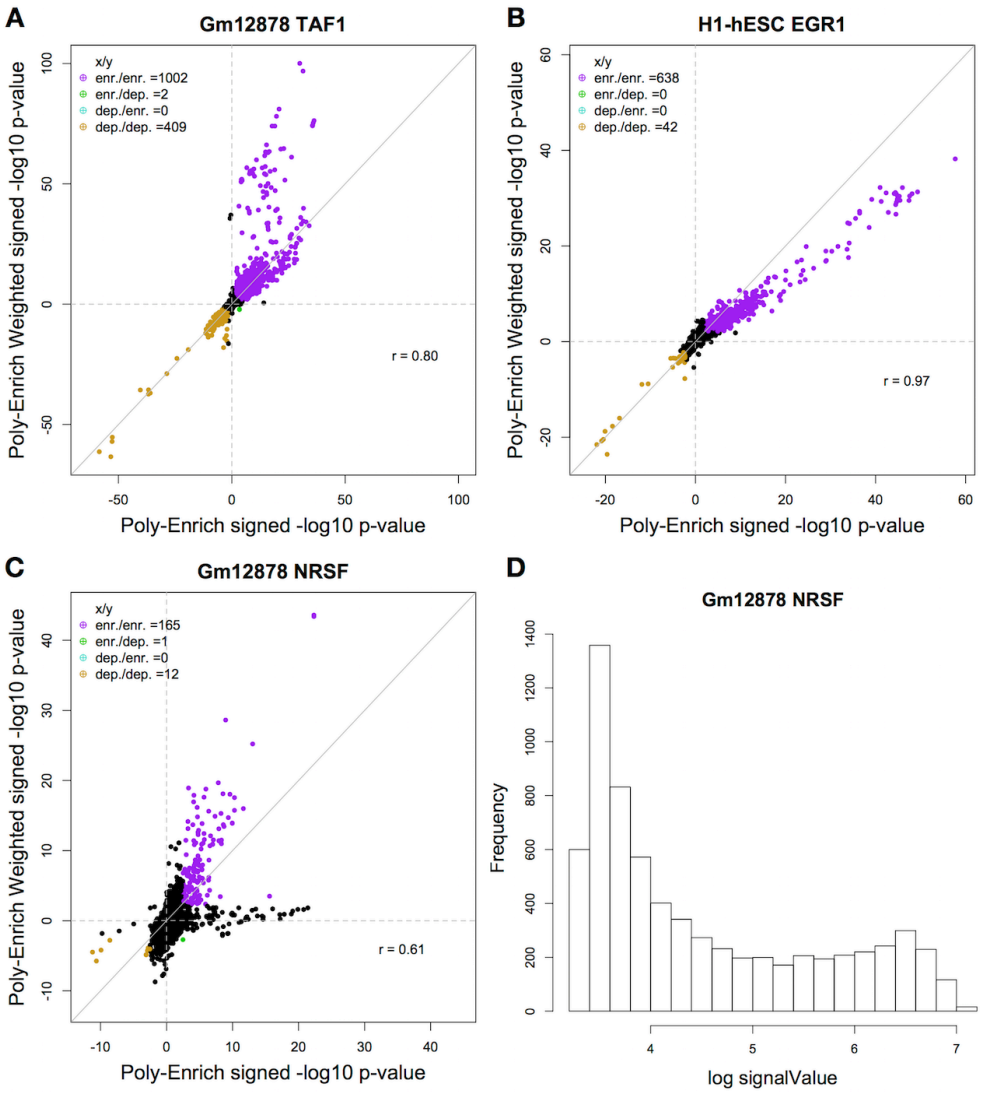


Figure 2.3 Comparison of GO term enrichment results between standard Poly-Enrich and its weighted version using signal values as weights.

Each point is a GO term's $-\log_{10}$ p-value of the two methods, signed positive for enriched, negative for depleted. (A) Using weighting results in more significant enrichment in many GO terms in the Gm12878 TAF1 ChIP-Seq experiment. (B) Using weighting results in slightly less significant enrichment in many GO terms in the H1-hESC EGR1 ChIP-Seq experiment. (C) Using weighting on the Gm12878 NRSF experiment results in several more significant GO terms as well as several less significant ones. (D) The histogram of (natural) log signal values from the NRSF experiment shows a bimodal pattern in the weights, suggesting that GO terms that are more significant with weighting than without may have genes that tend to have stronger bound peaks or vice versa.

Comparison of the count-based (Poly-Enrich) versus binary (ChIP-Enrich) model of enrichment

We next compared results from Poly-Enrich versus ChIP-Enrich on the same set of 90 ENCODE ChIP-seq datasets. Our initial hypothesis was that some experiments would be clearly modeled better by one method or the other (i.e. dependent on the transcription factor). However, our results strongly suggest that the optimal model for TF binding is more dependent on the gene set tested than the TF. This is visualized by a bifurcation in the significance levels of GO terms between the binary and count-based methods (Figure 2.4A), and suggests that a single transcription factor may regulate genes differently depending on the function of the gene. Thus, we sought to understand this further.

The binary model used by ChIP-Enrich assumes that a single binding event (i.e. a single genomic region) is sufficient for regulation, while the Poly-Enrich count-based model assumes that strength of regulation is incremental with the number of binding sites. Based on the results above, we asked what kinds of genes were more consistent with either of those assumptions. We use the true positive set of known TF-GO combinations mentioned earlier in the validation section. Observing the enrichment results using the 5kb locus definition for these true positive GO term-TF pairs, we used clustering to identify patterns of TFs and GO terms that are optimal with one of the methods. We found that the method that worked better was most often determined by the GO term (Figure 2.4B). For example, GO terms involving positive regulation of metabolic or biosynthetic processes tended to do better with Poly-Enrich except for those involving cell cycle, implying that related genes are regulated such that more binding sites increase regulation (Figure 2.4C,D). Conversely, GO terms related to “cell cycle” clustered together and displayed greater power with ChIP-Enrich, implying that related genes are possibly regulated with only one binding site and having more have little additional effect. Parallel results using the Nearest TSS locus definition were similar (Supplementary Figure 2.5).

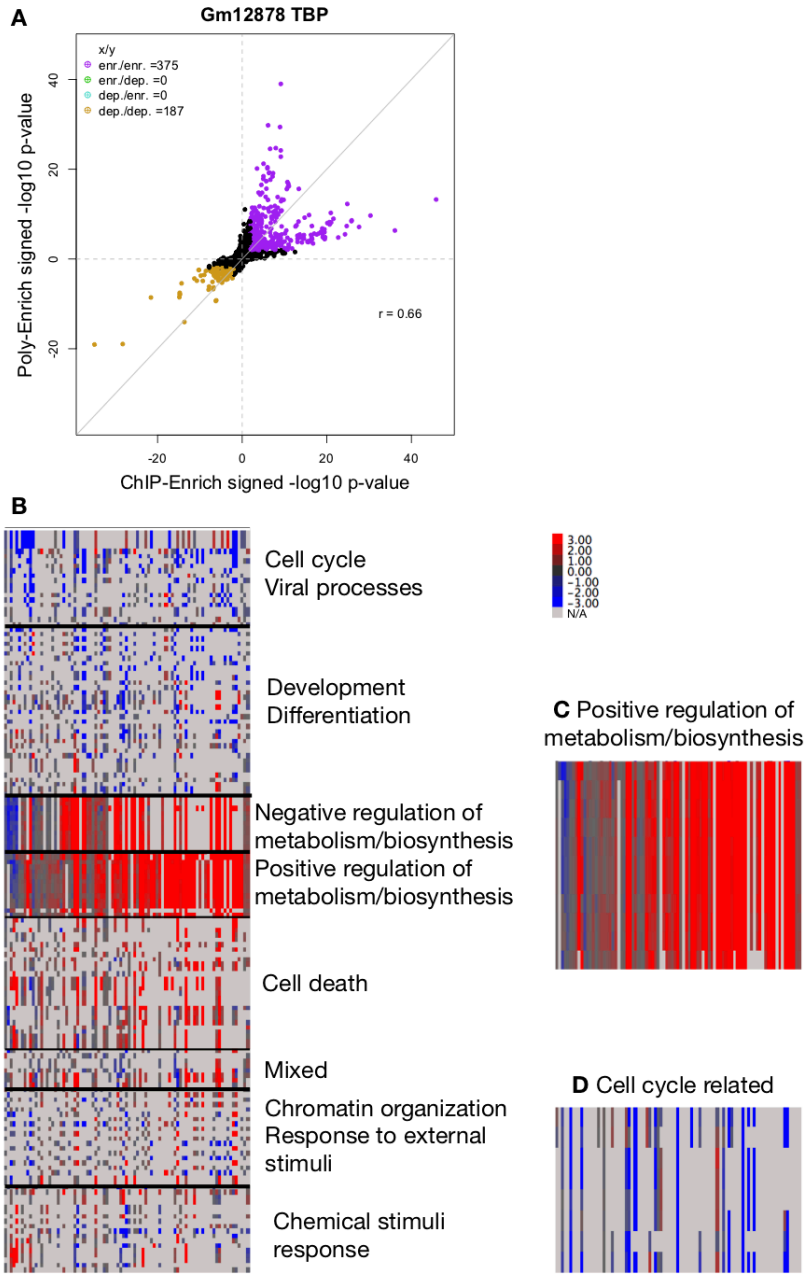


Figure 2.4 Comparisons of Poly-Enrich with ChIP-Enrich.

(A) Comparison of GO term significance levels between ChIP-Enrich and Poly-Enrich. Each point is the $-\log_{10}$ p-value of a GO term from the two methods, signed positive for enriched or negative for depleted. Several gene sets are much more significant using Poly-Enrich and several are much more significant using ChIP-Enrich. This split pattern is representative of 32% of the tested datasets. (B) Heatmap of $-\log_{10}$ p-value differences between Poly-Enrich and ChIP-Enrich for GO

terms and ChIP-seq experiments, where each row is a GO term and each column is a ChIP-seq experiment. Shown are GO terms where more than 15% of the experiments had a $-\log_{10}$ p-value difference of 2 or larger. Red indicates Poly-Enrich was more significant, and blue indicates ChIP-Enrich was more significant. Light grey indicates the transcription factor used in the experiment was not assigned to the GO term and is omitted in the clustering. Representative GO terms are shown for each cluster. (C) GO terms containing "positive regulation of metabolism/biosynthesis" are mostly red, indicating that a count score provides a more appropriate model. (D) GO terms related to cell cycle are mostly blue, indicating that a binary score provides a more appropriate model.

Poly-Enrich is recommended for experiments with a large number (>40k) of peaks, as we showed that ChIP-Enrich starts losing power at around 100ks of peaks (Supplementary Figure 2.4C). However, in many cases, the gene set, rather than the transcription factor, was a stronger determinant of the more appropriate method, we are not always able to recommend either Poly-Enrich or ChIP-Enrich for an entire experiment. We therefore developed a hybrid test that uses information from both ChIP-Enrich and Poly-Enrich.

Hybrid test

To obtain the best results across all types of GO terms and datasets, we developed a hybrid test that incorporates both the binary and count-based models. After performing the two models, the hybrid p-value of the two tests is defined as: $p_{hybrid} = 2\times\min(p_{CE}, p_{PE})$, where p_{CE} and p_{PE} are the p-values given by ChIP-Enrich and Poly-Enrich, respectively (Zhang et al. 2016). This is essentially a Bonferroni-adjusted p-value for two tests. This hybrid has been shown to be beneficial if the two tests are sufficiently different, but loses power and is conservative if the tests are identical or nearly identical (Zhang et al. 2016). While the hybrid test is not as powerful as the better method between ChIP-Enrich and Poly-Enrich, it is dramatically more powerful than using the worse method, making

it the optimal method to use across all GO terms (Figure 2.5). While this hybrid test currently only accommodates ChIP and Poly-Enrich, it can be extended to accommodate several additional gene set enrichment tests.

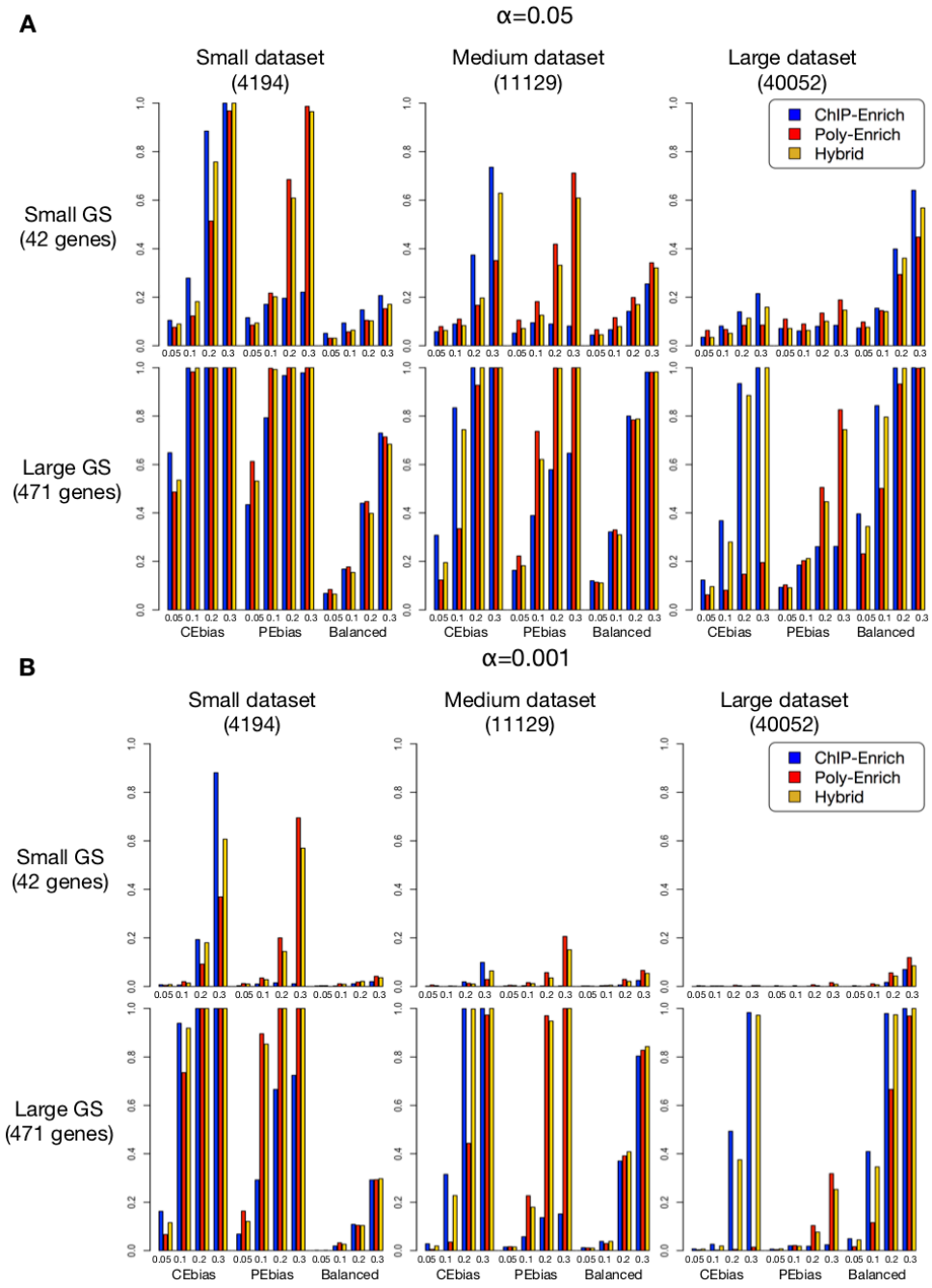


Figure 2.5 Statistical power comparisons for Poly-Enrich (red), ChIP-Enrich (blue), and the hybrid test (gold).

We compared datasets of three different sizes (i.e. number of peaks: small, medium, and large) and two gene set sizes (small and large GS), under two significance levels:

α = 0.05 (A) and 0.001 (B), and three different methods of simulated enrichment (CEbias: add peaks according to the regulatory assumptions of ChIP-Enrich, PEbias: add peaks mainly according to the assumptions of Poly-Enrich, Balanced: add peaks proportional to each gene's locus length). The values on the X-axis indicate the percent of extra peaks added to simulate enrichment; a higher value simulates stronger enrichment. The hybrid test is shown to have much more power than the wrong method, and only slightly less power than the correct method.

Identifying biological processes enriched with or depleted in repetitive element families using Poly-Enrich

ChIP-Enrich is unable to identify enriched gene sets in cases where nearly all genes have at least one assigned genomic region (Figure 2.1c). Thus, to further illustrate the utility of Poly-Enrich, we used it to test large families of repetitive element regions. We asked whether we could identify gene sets that are either enriched or depleted for certain types of repetitive elements. Significant enrichment of repetitive elements in the promoter regions of genes, for example, can sequester the transcription factors that inhibit activities at another transcription factor binding site or other regulatory motif (Liu et al. 2007). Some of these mobile elements remain active with new insertions having neutral, detrimental, or beneficial effects. Although repetitive element families have been well studied for over 30 years, little is yet known about the biological processes that they have adapted to help regulate or that they can easily disrupt and thus are negatively selected against (Brunner, Schimenti, and Duncan 1986). Using the database of human repetitive elements from the UCSC Table Browser (RepeatMasker 3.0) (Tarailo-Graovac and Chen 2009), we performed GSE testing on repetitive element families. Certain families of repetitive elements have over a million occurrences across the human genome, and thus virtually all genes have at least one nearby instance, making this an example where ChIP-Enrich performs poorly. Thus, in this situation, modeling the number of insertions per gene is critical to identify differences.

We examined two of the most abundant types of repetitive elements: the *Alu* and LINE1 (L1) elements, which make up an estimated 11% and 17% of the human genome, respectively (Roy-Engel et al. 2001, Lander et al. 2001). We also chose four

gene locus definitions: Nearest TSS, <5kb (promoter regions), >5kb (distal regions), and Intron. We tested GO Biological Processes, and used clustering to identify related groups of biological processes enriched with or depleted of the repetitive elements (Figure 2.6). We found that both Alu and L1 elements are enriched in centrosome-related GO terms, which validates that our approach identifies known associations (de Sotero-Caio et al.), and was only made possible with recent advancements in genome mapping near the centromeres (Aldrup-Macdonald and Sullivan 2014). For Alu elements, we also found strong enrichment in GO terms describing metabolic processes, most significantly "ATP metabolic process" and "rRNA metabolic process", especially in promoter regions, which is consistent with an analysis of *Alu* distribution in chromosomes 21 and 22 that showed *Alu* elements on these chromosomes were enriched in or near metabolism and signaling genes (Wanichnopparat et al. 2013). Conversely, Alu elements were sharply depleted in the promoter regions of many development and morphogenesis processes, with the strongest depletions in *cell fate commitment* and *connective tissue development*. Interestingly, depletions were also seen in the introns of genes in these gene sets, but not in regions >5kb upstream, suggesting the negative selection is limited to the regions that are more commonly regulatory.

Novel insertions of L1 elements into or near key genes are known to be associated with neurological diseases (Solyom and Kazazian 2012). Consistent with this, we found that all neuro-related GO terms in Figure were depleted for L1 (but not for all of Alu) (Supplementary Figure 2.7), which suggests that L1's evolutionarily have been selected against occurring in the regulatory regions of neurological genes; when they are inserted into the introns or promoters of these genes, the inserted elements may have an unacceptably high risk of causing disease.

In general, we observed that the significance of the distal upstream regions (>5kb locus definition) was lower than the other three locus definitions (with the exception of some enrichments for Alu elements) (Supplementary Table 2.3), implying that most repetitive element negative (or positive) selection has occurred

in the promoter regions or introns of genes. Alternatively, the gene distal enriched and depleted regions may be limited to a specific set of enhancer regions, the signal from which could have been diluted in our analysis. Interesting additional findings are that L1 elements are enriched in chemical stimulus detection processes such as *detection of chemical stimulus* and *sensory perception of chemical stimulus*, while Alu elements are depleted in the genes in these processes. Finally, both Alu and L1 elements are significantly depleted in genes involved in many processes related to development and morphogenesis.

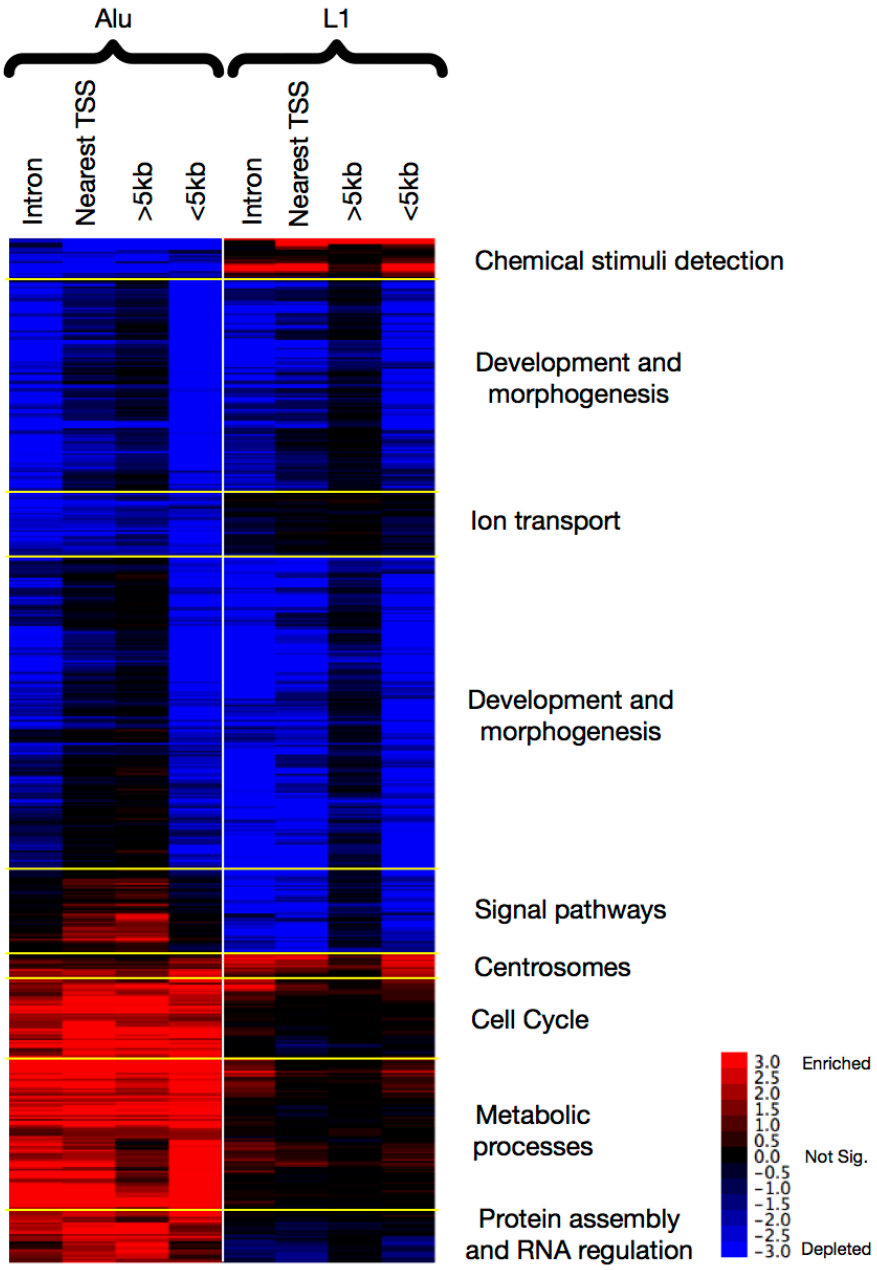


Figure 2.6 Gene Ontology terms enriched or depleted with common repetitive element families.

Shown are enrichment results using Poly-Enrich for the Alu (first four columns) and L1 (last four columns) repetitive element families using four different peak-to-gene assignments. Shown are signed $-\log_{10}$ FDR, where positive values (red) indicate enrichment and negative values (blue) indicate depletion. Only GO terms that were significant for at least 3 columns at the FDR = 0.05 level are displayed. We identified nine clusters of GO terms with similar enrichment patterns. Representative GO terms are used to label each cluster.

Identifying an optimal enhancer locus definition

One shortcoming of our current methods (as well as current alternatives) is that they rely on associating each genomic region with the nearest gene(s). However, it is estimated that 79-95% of DNAse I hypersensitive sites, markers for enhancer regions, actually regulate a different, distal target gene (Boyle et al. 2008, Thurman et al. 2012). To improve upon the nearest gene approach, we sought to compare a large number of enhancer locus definitions. We generated a set of enhancer locus definitions that identify and assign enhancer regions to their appropriate target genes, as was recently introduced by Chicco D, et al (Chicco et al. 2019), so peaks in enhancer regions will be correctly assigned and false positive peaks in nonfunctional intergenic regions will be filtered out. Details on the generated locus definitions can be seen in the supplementary material (Qin et al.).

We used the score test approximation of Poly-Enrich for preliminary results of our set of 1,860 enhancer locus definitions with over 87 ChIP-seq experiments from ENCODE (excluding Pol2). We then used the true positive set as defined above to evaluate each enhancer definition with an F1 score, or the harmonic mean of precision and recall (Sasaki 2007), based on the average across the 34 TFs. After filtering down to the top 10 locus definitions by F1 score, we used standard Poly-Enrich to finalize the ranking. We found that these locus definitions significantly outperform the locus definition where each peak is naïvely assigned to the nearest gene TSS (Figure 2.7). Currently, the best performing locus definition has enhancer locations from DNAse hypersensitive sites (DHS) from 125 cell types processed by ENCODE and other DHS's that are within 500kb from other DHS's in 32 cell types (Thurman et al. 2012). The enhancer targets are the union of ChIA-PET interactions (Li et al. 2017), DNAse signal correlations (Thurman et al. 2012), and links from the FANTOM5 dataset (Lizio et al. 2015). This enhancer locus definition is available for genomes *hg19* and *hg38,* and work is ongoing to implement *mm9* and *mm10*. As research for other species' enhancer regions and targets are currently less

established, making enhancer locus definitions for those species likely results in suboptimal analyses.

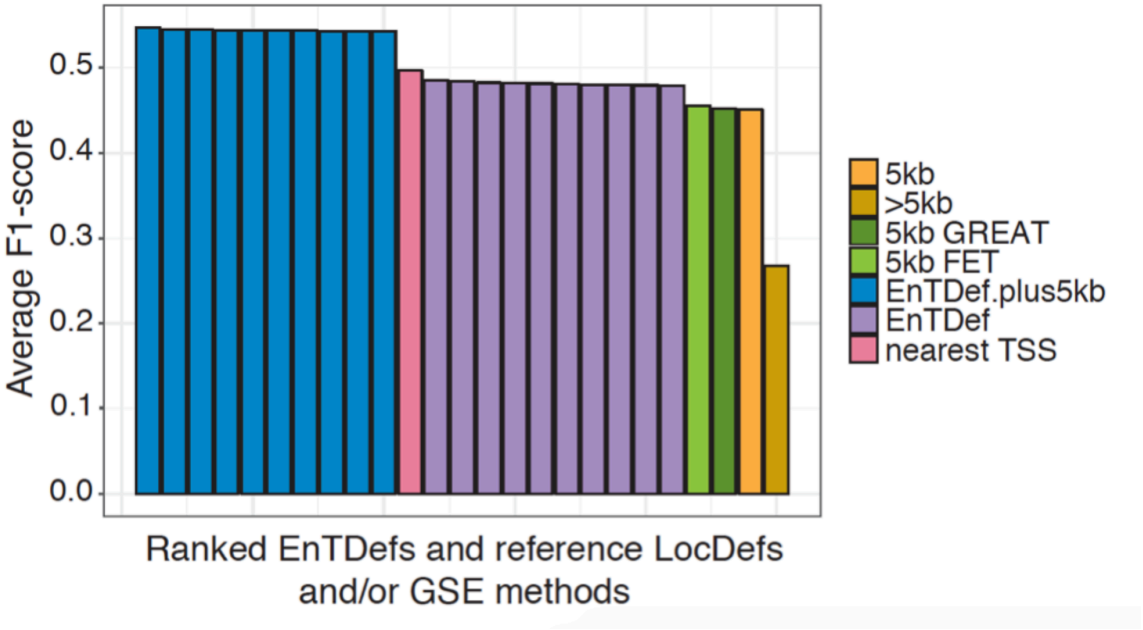


Figure 2.7 Ranked enhancer locus definitions by average F1-score across the 87 ChIP-seq experiments.

The top 10 enhancer locus definitions (blue) perform consistently better than the nearest TSS locus definition (pink). When comparing the top 10 locus definitions without the promoter region (purple) to the locus definition of nearest TSS excluding the promoter region (labeled ">5kb" in mustard yellow), we see a large improvement in F1 score, showing the new enhancer-gene targets are an improvement to naïvely assigning peaks to their nearest gene TSS.

Availability, usage, and updates

Poly-Enrich is available in the *chipenrich* Bioconductor package and as a web interface at http://chip-enrich.med.umich.edu. Several additional gene set databases and gene locus definitions (see *Methods* for details) have been added since our original publication (see http://chip-enrich.med.umich.edu/data/ChipenrichMethods.pdf).

To perform GSE analysis with either our Bioconductor package or web version, the user first needs a file of genomic regions, which may be a narrowPeak, BED, or text

file with chromosome, start, and end positions for each region. The user then selects a species, one or more gene set databases, a gene locus definition, and the test method (ChIP-Enrich, Poly-Enrich, Hybrid, or Fisher's exact test). Optionally, the user can upload a custom/user-defined list of gene sets and/or gene locus definition. For narrow genomic regions (≤ 2 – 3 kb), we recommend using the Poly-Enrich method for sets of more than 100,000 regions, and the Hybrid method for sets of regions with fewer than this. For broad genomic regions (> 2 – 3 kb), we still recommend the Broad-Enrich method (Supplementary Figure 2.8). The user can then also choose to weight the genomic regions based on a score of their choice, and apply a number of other options, such as adjustment for read mappability (recommended for read lengths < 50 bp).

The enrichment function outputs five files:

- *opts*: The options that the user input into the function.
- *peaks*: A peak-level summary showing the peak-to-gene assignment for each peak.
- *peaks-per-gene*: A gene-level summary showing gene locus lengths and the number of peaks assigned to each gene.
- *results*: The results of the GSE tests. Lists the tested gene sets along with their descriptions, the test effect, odds ratio, enrichment status, p-value, and FDR. Also included is the list of Entrez gene IDs with contributing signal for each enrichment test.
- *qcplot*: A diagnostic plot of the gene locus lengths with a fitted smoothing spline.

The R code used to generate analysis and figures can be found at: https://github.com/sartorlab/polyenrich

Discussion

Gene set enrichment testing methods for genomic regions should take into account the differing properties of the input datasets, including the widths and number of

genomic regions, and where they tend to occur relative to genes. However, no single method is appropriate for all types, and therefore no single GSE method should be recommended for all sets of genomic regions. Although our previously developed ChIP-Enrich method for gene set enrichment with genomic regions performs well for most transcription factor ChIP-seq datasets (Welch et al. 2014), above we described common situations where it does not. Such cases include when nearly all genes are assigned at least one genomic region, and when the strength or likelihood of regulation increases incrementally with the number of genomic regions. As an example, the transcription factor NF-kappaB is known to regulate the gene NFKBIA by binding to a few or even many motif positions in the promoter (Giorgetti et al. 2010), with gene expression correlated with the number of bound factors. Thus, motivated by specific examples of regulatory mechanisms, we developed Poly-Enrich, a method that models the number of regions per gene, empirically adjusts for each gene's locus length, and takes into account variability among genes in each gene set. Poly-Enrich is also flexible, in that it easily allows for weighting of each genomic region by any score of interest. We used the example of weighting by peak strength, but other examples include weighting by SNP significance in a GWAS analysis, by the inverse distance to a gene, or by the probability that the region is in an open chromatin region in a particular cell type.

We showed that our count-based method, Poly-Enrich, is optimal when almost all genes are assigned a peak. In comparing when each test is most appropriate for typically sized ChIP-seq datasets, we discovered that the optimal test is mostly dependent on the gene set, rather than the transcription factor being studied. Because in many cases we could not recommend a single best method to test all gene sets for an experiment, we developed and implemented a hybrid test that uses information from both methods and performs better than either test across GO terms for most datasets. However, as noted in the Results, specific situations exist when one particular method is optimal, and we therefore have provided specific recommendations to our users in choosing the most appropriate method. Our hybrid test is currently a very simple approach, but as both ChIP-Enrich and Poly-

Enrich are based on well-known GLM link functions, it is possible to use a theoretical correlation to give a less conservative but still well-controlled hybrid p-value.

When applying Poly-Enrich to repetitive element families, we both reconfirmed known associations and also identified novel findings. Poly-Enrich confirmed that Alu elements are overrepresented in the promoters of metabolism genes and signaling by finding enrichment for related GO terms. Additionally, we know that L1 insertions into or near certain neurological-related genes are associated with neurological diseases (Thomas, Paquola, and Muotri 2012). Indeed, we found that L1 is depleted in neuro-related process genes, implying there is natural selection against L1 elements inserting in the regulatory regions of these genes. We also found that there is little enrichment or depletion in the distal regulatory regions of genes, suggesting that repetitive elements may not have as large of an effect there due to mitigated regulatory activity at larger distances from transcription start sites. We also detected novel associations between repetitive element families and biological pathways. Both Alu and L1 elements were significantly depleted in development and morphogenesis-related gene sets, such as *connective tissue development* and *skeletal system morphogenesis*, suggesting that it is critical to have developmental regulatory regions for several different development systems free from potentially disruptive repetitive elements during early growth.

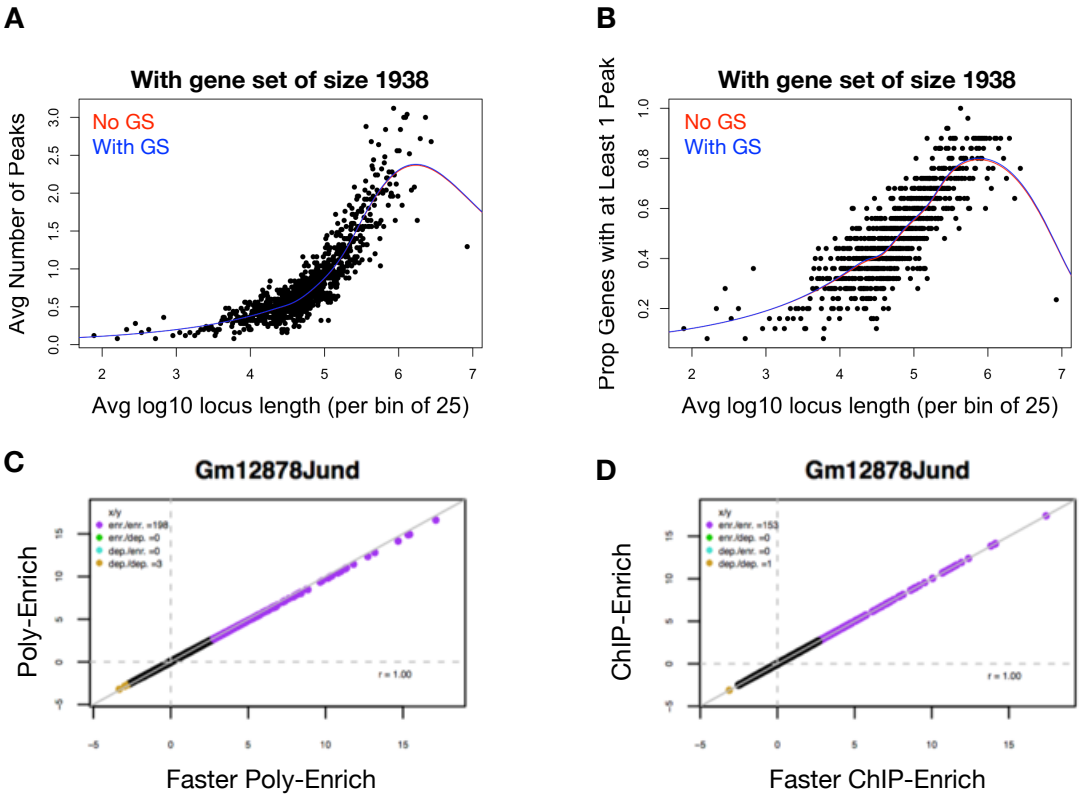
We also used Poly-Enrich to find an optimal enhancer locus definition that is superior to naïvely assigning each peak to the gene with the nearest TSS. This will greatly improve analysis of TFs that tend to bind in distal regulatory regions, as the peaks will be much more likely to be assigned to the gene it is truly regulating. However, these locus definitions were defined by currently available research, and there may be more enhancer regions or enhancer-gene links discovered in the future that may provide a more accurate enhancer locus definitions.

Supplementary Material for Chapter 2

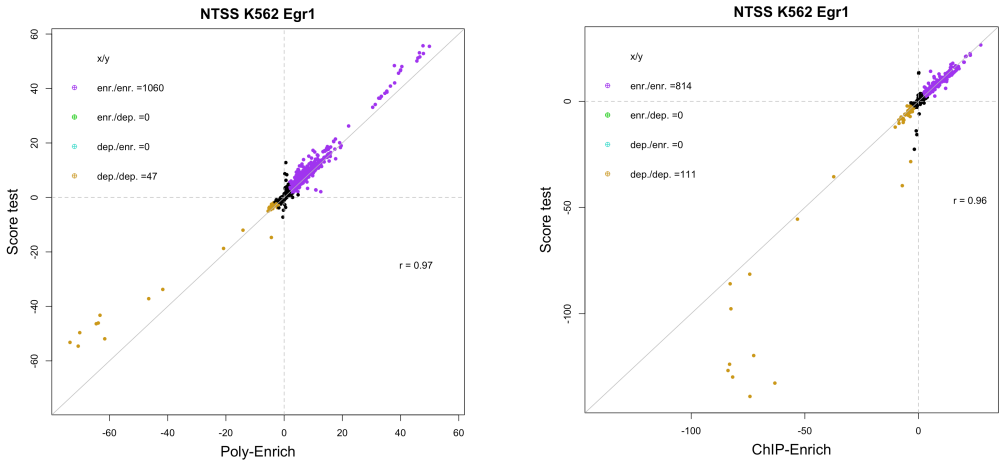
Generation of enhancer locus definitions

For each enhancer locus definition, there is two parts: the definition of all the enhancer regions on the genome, and enhancer-gene target links for each region. The enhancer regions were identified from four data sources: ENCODE ChromHMM USCS tracks (Ernst and Kellis 2012), DNAse hypersensitive sites from ENCODE (Thurman et al. 2012), Cap Analysis Gene Expression experiment-derived enhancers from the FANTOM5 project (Lizio et al. 2015), and distal or non-promoter DNAse hypersensitive sites within 500kb of a correlated promoter from a publication from Thurman et al (Thurman et al. 2012). The enhancer-gene target links were defined from four methods: ChIA-PET2 interactions (Li et al. 2017) from two ChIA-PET datasets (Dowen et al. 2014, Tang et al. 2015), DNAse-signal correlation identified by Thurman et al (Thurman et al. 2012), enhancer-gene links from the FANTOM5 project (Lizio et al. 2015), and loop boundaries with CTCF motif (Rao et al. 2014). To improve gene coverage, we had the choice of extending each enhancer region by 1kb, as well as assigning the remaining unassigned enhancer regions or all unassigned peaks to their nearest gene TSS. In total, we made 1,860 enhancer locus definitions.

Supplementary Figures for Chapter 2



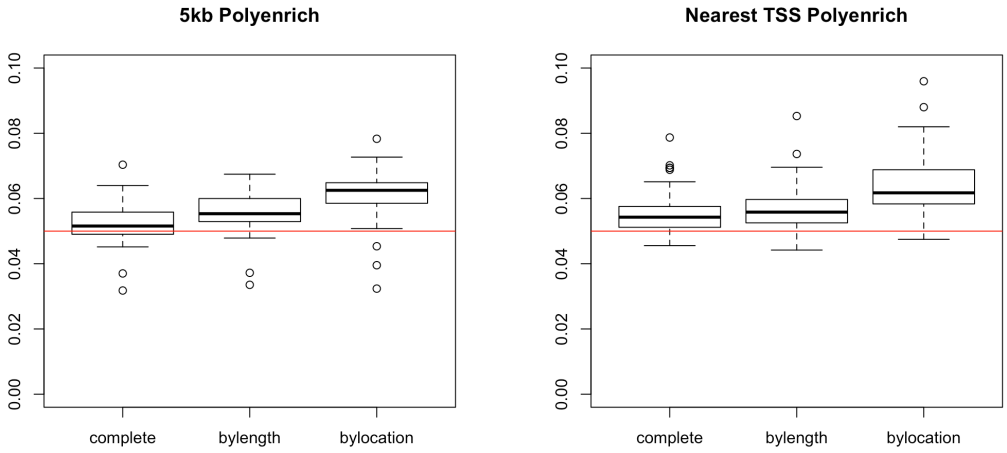
Supplementary Figure 2.1: (A) A comparison of the Poly-Enrich spline estimates with and without the Gene Set (GS) covariate for a gene set with 1938 genes (out of ~20000). There is very little difference despite this gene set being larger than 97% of all gene sets (we chose a large gene set because it would have a greater effect on the spline than a small gene set). For visualization only, each point is the average number of peaks assigned to each gene within sequential bins of 25 genes. The large locus length outlier is from genes near the ends of chromosomes that tend to have very low binding activity. (B) Similar to A, except using the proportion of peaks with at least one peak as the measure, as used for ChIP-Enrich. Each point is the average number of peaks within sequential bins of 25 genes. (C) Signed -$\log_{10}$ P-value comparisons between Poly-Enrich and its faster counterpart which uses a spline approximation. (D) Signed -$\log_{10}$ P-value comparisons between ChIP-Enrich and its faster counterpart that uses a spline approximation. As the gene set-specific spline is nearly identical to the approximated spline, using the spline approximation does not change the results by much, with an almost perfect identity trend with $r^2 = 1.00$.



Supplementary Figure 2.2: Score test comparisons for Poly-Enrich and ChIP-Enrich. All GO terms are generally concordant with the score test being slightly more conservative, but the depleted GO terms tend to deviate more.

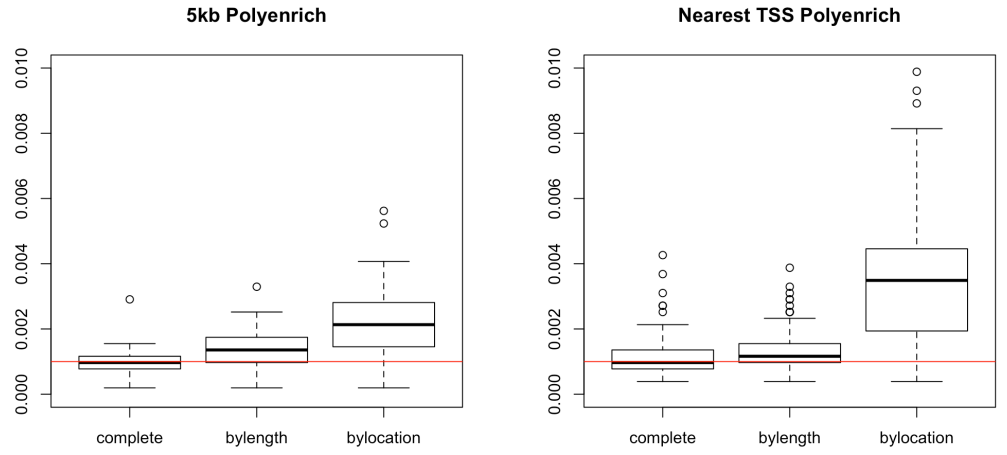
A

$\alpha = 0.05$

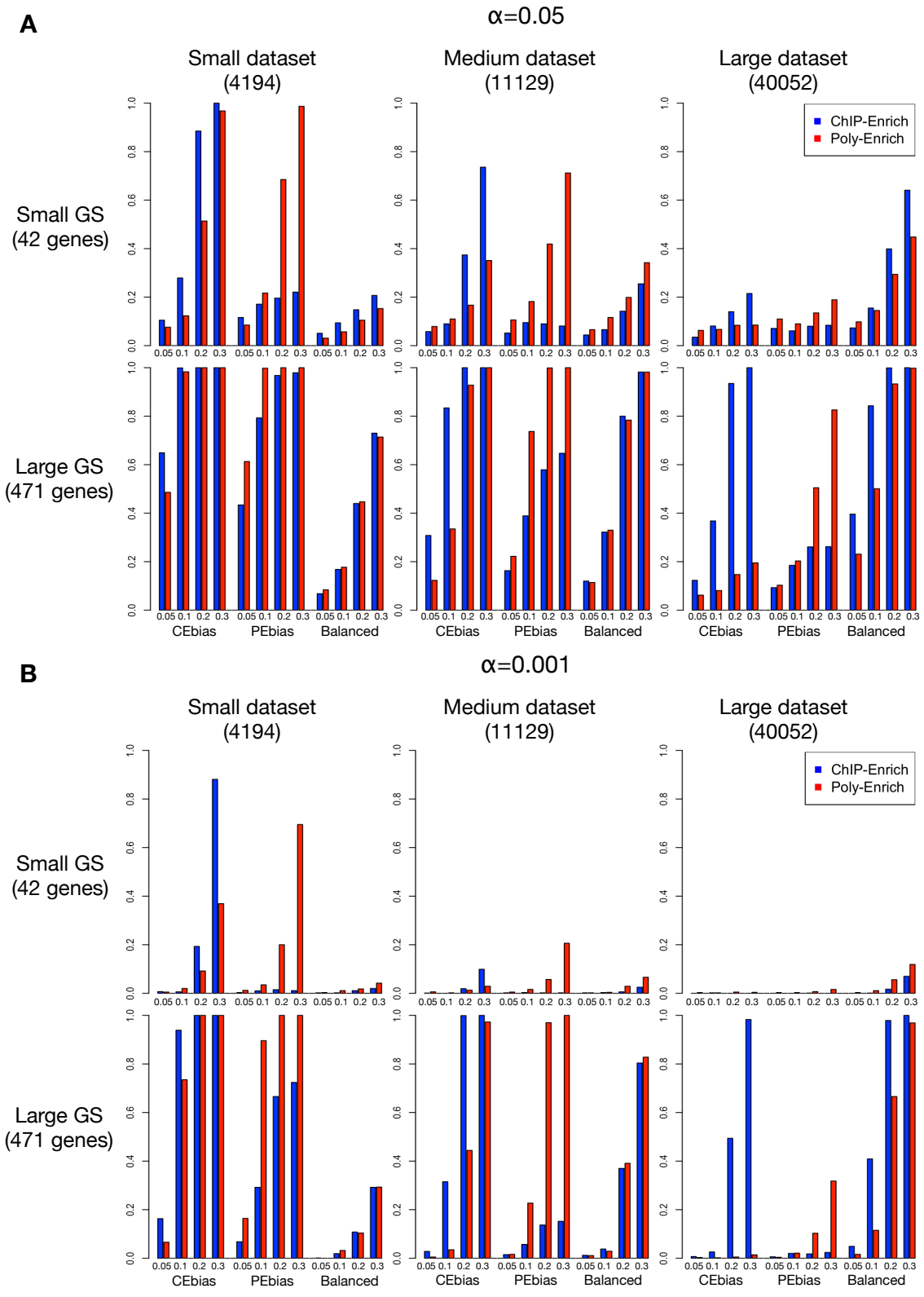


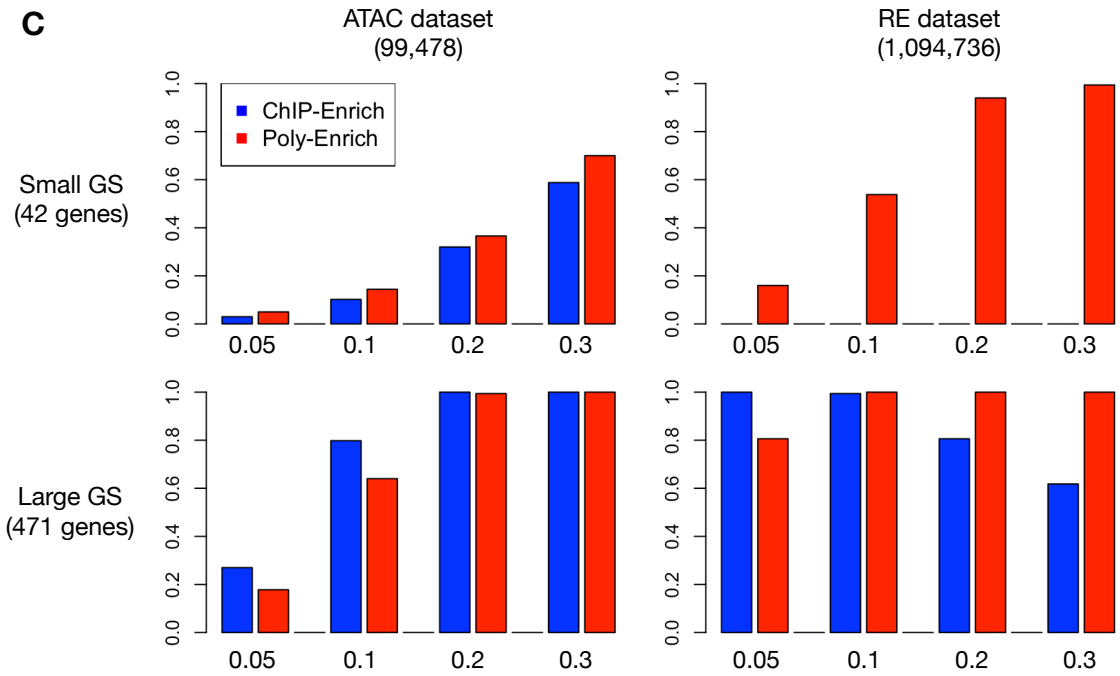
B

$\alpha = 0.001$

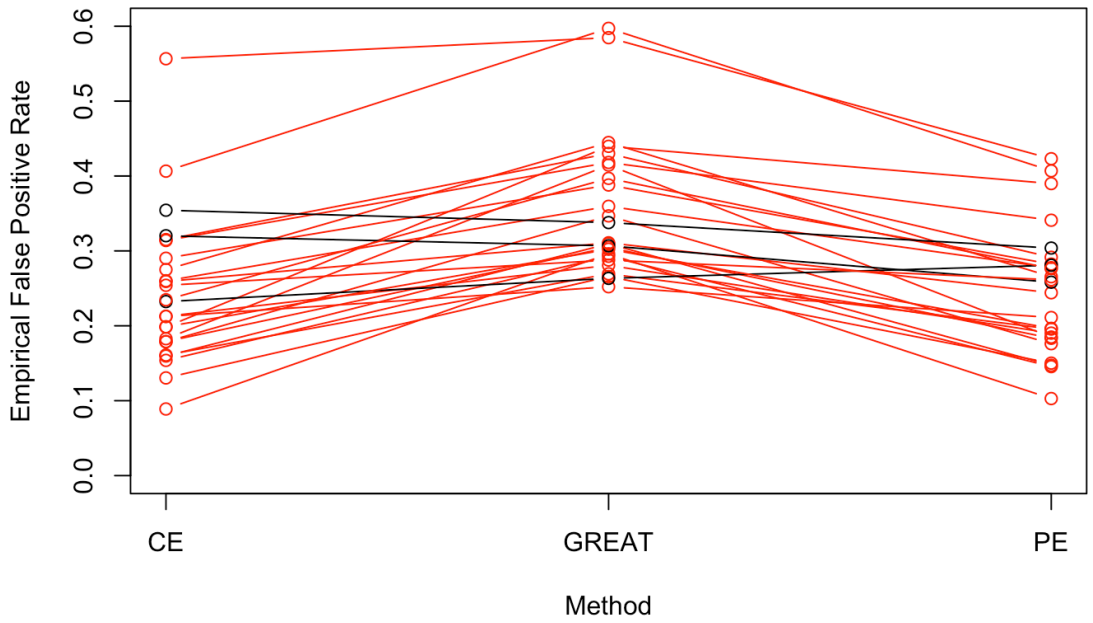


Supplementary Figure 2.3: Poly-Enrich Type I error rate plots using the <5kb and Nearest TSS gene locus definitions for levels 0.05 (A) and 0.001 (B). Shown are the proportion of significant gene sets out of 50,150 (5,015 GO terms × 10 permutations) randomized gene sets for each of 90 ENCODE ChIP-Seq datasets. Type I error rates are acceptable at the 0.05 and 0.001 level. There is some inflation at the 0.001 level for the *bylocation* randomization, which is caused by several related genes clustering near each other, causing consistent significance in some GO terms. (Supplementary Table 2.2).

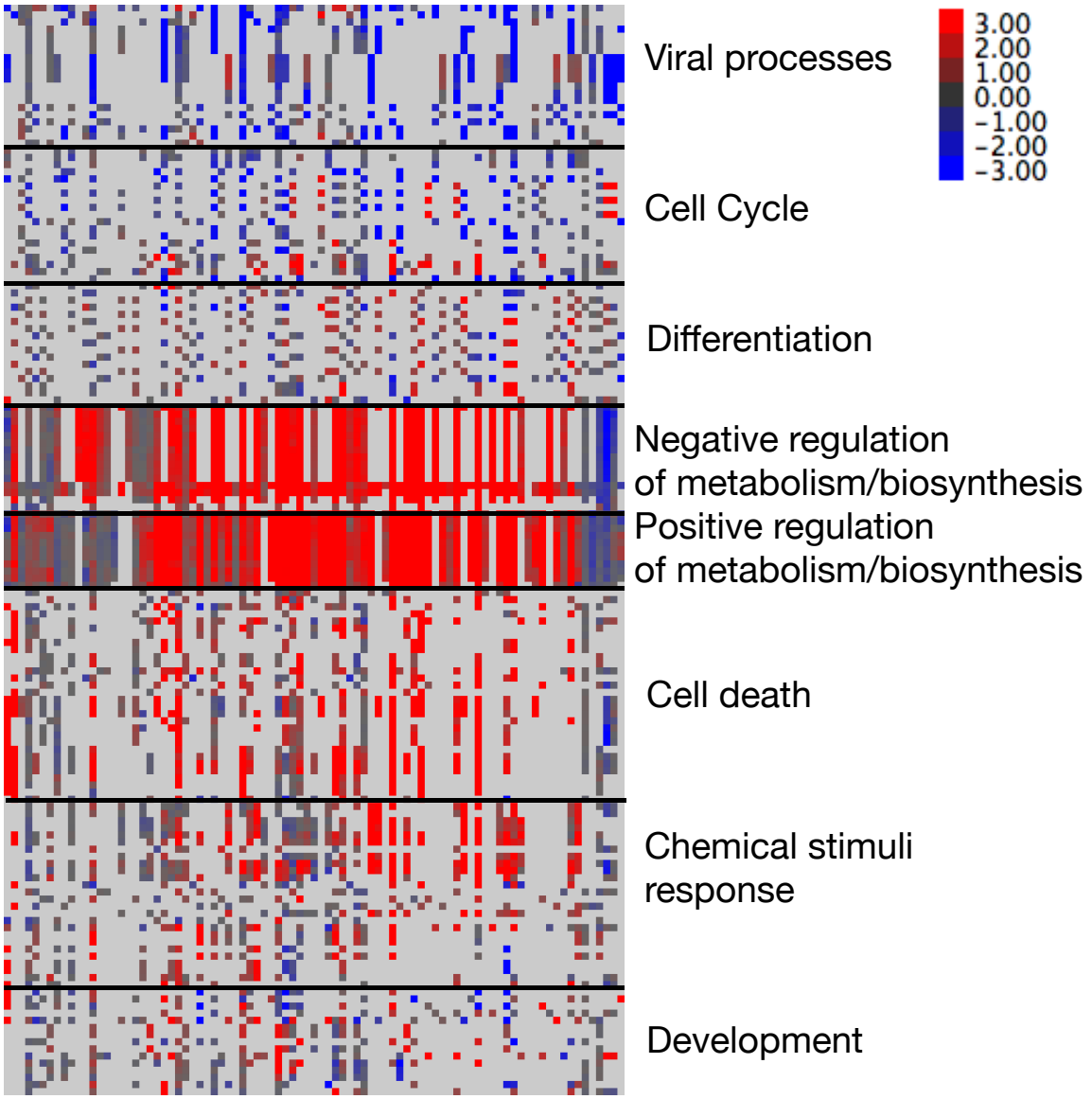

A
α=0.05
Small dataset
(4194)
Medium dataset
(11129)
Large dataset
(40052)
ChIP-Enrich
Poly-Enrich
Small GS
(42 genes)
Large GS
(471 genes)
CEbias
PEbias
Balanced
B
α=0.001
Small dataset
(4194)
Medium dataset
(11129)
Large dataset
(40052)
ChIP-Enrich
Poly-Enrich
Small GS
(42 genes)
Large GS
(471 genes)
CEbias
PEbias
Balanced



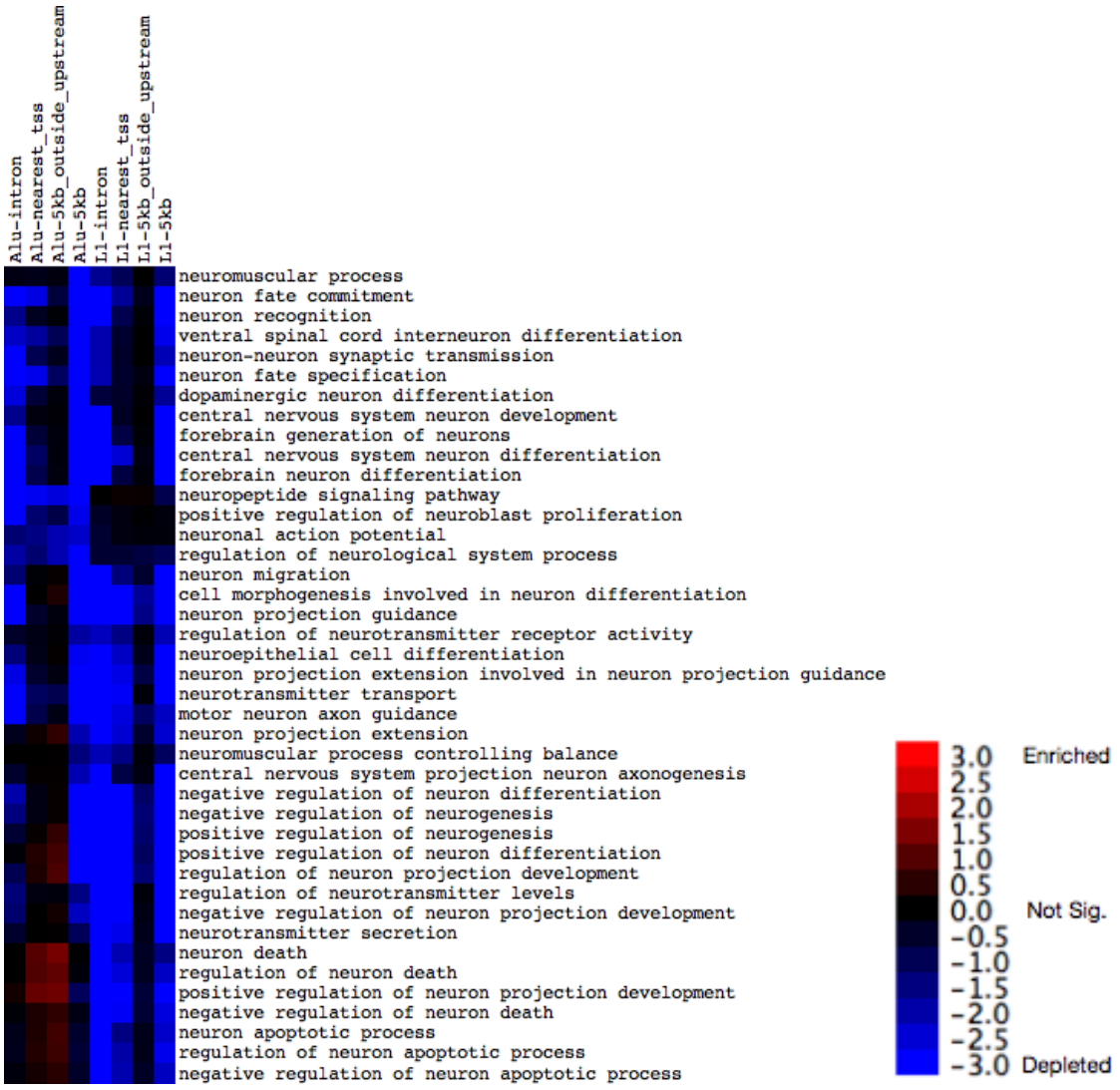
Supplementary Figure 2.4: Statistical power comparisons between ChIP-Enrich (blue) and Poly-Enrich (red) for datasets with three different sizes (i.e. number of peaks: small, medium, and large) and two gene set sizes (small and large GS), under two significance levels: α = 0.05 (A) and 0.001 (B), and three different methods of simulated enrichment (CEbias: add peaks according to the regulatory assumptions of ChIP-Enrich, PEbias: add peaks mainly according to the assumptions of Poly-Enrich, Balanced: adding peaks in proportion to each gene's locus length). The values on the X-axis indicate the percent of extra peaks added to simulate enrichment; a higher value simulates stronger enrichment. A stricter significance level results in less power and a larger gene set results in more power. Simulations on larger datasets artificially reduced power because randomized larger datasets include more noise. However, in real experiments, we expect larger datasets (more peaks) to be more powerful, since the majority of the peaks in the dataset are not expected to be noise. (C) Additional power comparisons using the Balanced simulations on larger data sets (ATAC-seq and repetitive elements) showing ChIP-Enrich can still do reasonably well at around 100k peaks, but starts having trouble on very large datasets where most (~81% in largest dataset) genes are assigned a peak.



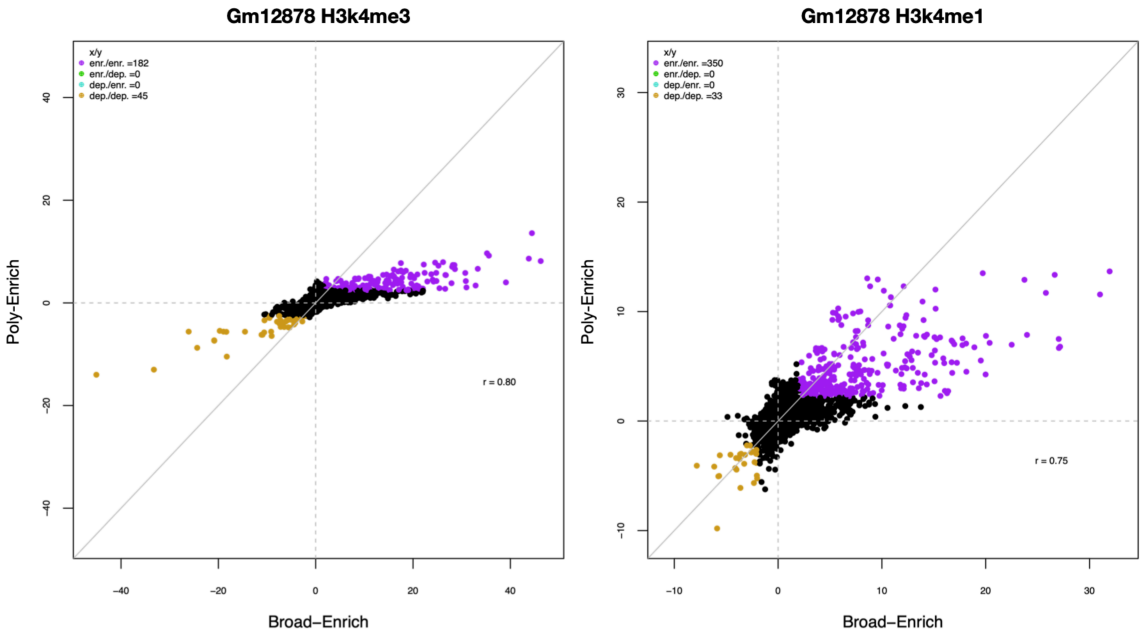
Supplementary Figure 2.5: Estimated false positive rate comparisons between ChIP-Enrich (CE), GREAT, and Poly-Enrich (PE) for 25 TFs using the true positive and negative GO terms as determined by GO.db (see Methods). Each TF's results are connected by lines, with the TFs where GREAT has a higher FPR than both methods colored in red. GREAT has higher FPR than both methods for 22 TFs. The overall high FPR is likely an artifact of not having a perfect gold standard; the estimated FPR provides an upper bound to the true FPR.



Supplementary Figure 2.6: Comparison of GO term significance levels between ChIP-Enrich and Poly-Enrich using the *nearest TSS* locus definition. Shown is a clustered heatmap of $-\log_{10}$ p-value differences between Poly-Enrich and ChIP-Enrich for GO terms and experiments, where each row is a GO term and each column is a ChIP-seq experiment. Shown are GO terms where more than 15% of the experiments had a $-\log_{10}$ p-value difference of 2 or larger. Red indicates Poly-Enrich was more significant, and blue indicates ChIP-Enrich was more significant. Light grey indicates the transcription factor used in the experiment was not assigned to the GO term and is omitted in the clustering. Representative GO terms are used to label each cluster.



Supplementary Figure 2.7: Neuro-related enrichment results using Poly-Enrich for Alu (first four columns) and L1 (last four columns) repetitive element families using four different peak-to-gene assignments. Shown are signed –log10 FDR, where positive values (red) indicate enrichment and negative values (blue) indicate depletion. GO terms containing "neuro" are almost all significantly depleted.



Supplementary Figure 2.8: -$\log_{10}$ p-value comparisons between Broad-Enrich and Poly-Enrich for two histone modification ChIP-seq peaks, which have broad peaks. Broad-Enrich has much more powerful signals than Poly-Enrich.

Supplementary Tables for Chapter 2

Supplementary tables can be found in my Github dissertation repository at https://github.com/leetaiyi/Dissertation

Supplementary Table 2.1: List of all 90 ENCODE datasets used for gene set enrichment testing. Downloaded from ENCODE Analysis data at UCSC: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/

Supplmentary Table 2.2: (A) The top 20 GO terms from 4 bylocation randomization trials of the Gm12878 ETS experiment with their p-values and genes with contributing signal. Immediately apparent is that there are several of the same genes in these GO terms. (B) The top 30 most common genes in the top 20 GO terms. One large cluster of genes in Chromosome 6 are the histone cluster 1 family members, which would likely be randomized together in the same bin in the bylocation algorithm, causing p-value inflation for GO terms with these genes. The H1hesc TBP experiment had similar findings (not shown).

Supplementary Table 2.3: Results of repetitive elements Poly-Enrich analysis. Complete table of signed $-\log_{10}$ p-values for all combinations of GO terms, locus definitions, and type of repeated elements. Column names are in the format of [repeated element]-[locus definition].

CHAPTER 3 ProxReg: Testing Proximity to Transcription Start Sites and Enhancers Complements Gene Set Enrichment Testing

A paper covering most of material in this chapter is in review at Frontiers in Genetics, with myself as first author.

Introduction

Cell development and differentiation depend on complex gene expression patterns which are precisely and spatiotemporally controlled. The complex process of gene regulation involves many different mechanisms, including regulation of transcription (Berger 2007, Deaton and Bird 2011), post-transcriptional regulation (Roundtree et al. 2017), and regulation of translation (Sonenberg and Hinnebusch 2009). Transcription is the first step to decode the genetic information from DNA to functional elements, and this process is regulated by many *cis*-regulatory elements across the genome (Wittkopp and Kalay 2011). *Cis*-regulatory elements include promoters, enhancers, silencers, and insulators, with promoters and enhancers being two important ones that can initiate transcription and are the most well-studied (Andersson 2015). Both promoters and enhancers are regions of DNA sequences that typically are a few hundred base pairs in length (Nguyen et al. 2016). Promoters are usually located immediately upstream of the transcription start sites on the 5' end of target genes (Sanyal et al. 2012) and recruit transcription factors and RNA polymerase II (RNAPII) to instruct the direction and initiation of transcription (Schoenfelder and Fraser 2019). Conversely, enhancers can be located upstream, downstream, or in the intron of the target gene or another unrelated gene

(Shlyueva, Stampfel, and Stark 2014) and bound by transcription factors (TFs) and cofactors to activate or increase the transcription rate of their target genes (Li, Notani, and Rosenfeld 2016). The protein sequences and regulatory motifs of many transcription factors are well conserved across living organisms, indicating that genome-wide gene regulatory mechanisms have important conserved properties (Lambert et al. 2018). However, some transcription factors such as ESR1 bind to different sets of target genes in a cell type specific manner (Gertz et al. 2012), resulting in complex and dynamic transcription factor regulatory programs. Thus, deciphering the rules of transcription factor binding events is a key step to understanding gene expression patterns and associated biological pathways.

A diverse collection of sequence-based approaches exist to probe the gene regulome (Pinsach-Abuin 2016). For instance, ChIP-seq can provide genome-wide information about gene regulation for specific transcription factors or chromatin marks by identifying thousands of genomic regions (i.e. peaks, which we will refer to for simplicity) across the genome (Schmidt et al. 2009). ATAC-seq and copy number variation (CNV) sequencing are also popular for studying genome-wide regulation (Buenrostro et al. 2015, Xie and Tammi 2009). Through the aforementioned sequencing data, we can identify significant peaks that were bound by a particular TF or modified chromatin mark (ChIP-seq), open chromatin regions (ATAC-seq), or regions with a copy number variation. We can further infer their underlying regulatory functions by associating the identified regions with target genes, whether predicted or verified. Since biological processes involve many genes and pathways, gene-centered analysis on regulome data may not be as informative as Gene Set Enrichment (GSE) testing (Subramanian et al. 2005).

Most GSE methods were developed for gene expression data, do not adjust for the varying lengths of genes or regulatory space between them, and thus are not generally appropriate for GSE testing with large sets of peaks. However, several GSE methods have been developed to specifically test sets of peaks, including GREAT (McLean et al. 2010), ChIP-Enrich (Welch et al. 2014), Broad-Enrich (Cavalcante et al. 2014), and Poly-Enrich (Lee et al. 2018). Among these, Poly-Enrich is the only method that counts genomic regions (which we will refer to as peaks for simplicity)

for each gene, adjusts for the varying lengths of genes and regulatory space between them, and provides a flexible approach with the ability to assign weights to peaks. Current methods for GSE testing of peaks focus mainly on the relationship between peaks and TSSs (promoters). However, although some transcription factors (e.g. E2F1 (Ertosun, Hapil, and Osman Nidai 2016) preferentially bind to promoters, others (e.g. FOXA1 (Pristerà et al. 2015) tend to bind enhancers, while still other TFs bind to both enhancers and promoters depending on context (e.g. master regulators, such as Serum response factor). Therefore, it is of great interest to know the patterns of transcription factor binding with respect to promoters and enhancers of the target genes and pathways. Although GREAT (McLean et al. 2010), ChIP-Enrich (Welch et al. 2014) and Poly-Enrich (Lee et al. 2018) incorporate distal binding events in their GSE testing, no method has been established for answering the question of whether a TF is binding closer to transcriptions start sites (TSSs), near enhancers, both, or neither for a specific gene set. Other methods such as ChIPseeker (Yu, Wang, and He 2015) and Seq2pathway (Wang, Cunningham, and Yang 2015) also perform GSE testing for genomic regions. Different from previous GSE testing methods that assign peaks to nearest TSS, ChIPseeker applies a max distance cutoff for assigning peaks to genes. Seq2pathway incorporates the significance of each genomic region and both coding and non-coding regions in GSE testing. Methods such as Cistrome-GO (Li et al. 2019) and TREG (Chen et al. 2013) incorporate the distance between ChIP-seq peaks and a gene's TSS into the GSE testing itself. Cistrome-GO integrates the peak distance to TSS and the peak number together to estimate the gene regulation potential. TREG collects the peak distances to a gene's TSS within a 2Mb window around each TSS into the gene set enrichment test. However, since these methods embed the information about binding proximity to a TSS within the test itself, it is difficult for the user to interpret the results with respect to this information, or separate the effect of proximity from that of enrichment. No method, to our knowledge, incorporates enhancer proximity.
Here, we propose a new method, Proximity Regulation (ProxReg) to address this shortcoming of current methods. By measuring the distance between each peak and the closest TSS (or enhancer) and then performing a modified two-sided Wilcoxon

rank-sum test, we test whether the peaks in a gene set are significantly closer to TSSs or enhancers than expected by chance. Our method, in combination with a GSE test, is able to provide additional evidence that a pathway is truly enriched and information on the regulatory mechanism for that enrichment. After validating the Type I error rate of our method, we test ProxReg by applying it, in combination with Poly-Enrich (implemented in the *chipenrich* Bioconductor package) to 90 ENCODE ChIP-seq datasets (Sloan et al. 2016), including 35 TFs. In many cases, this led to a significant improvement in the ability to pinpoint the known biological processes in which a TF functions. In summary, we show the power and benefits of ProxReg, which is available in 5 species (fruit fly, zebrafish, mouse, rat, and human) for promoters and in human for enhancers, to complement GSE testing of large sets of peaks.

Materials and Methods

Datasets used

We used a total of 90 human ChIP-seq datasets from the Encyclopedia of DNA Elements (ENCODE) at University of California, Santa Cruz (Qu and Fang 2013, Consortium 2004, Sloan et al. 2016) that consists of 35 transcription factors over the three tier 1 cell lines (embryonic stem cells [H1-hESC], B-Lymphocyte [GM12878], and myelogenous leukemia cell [K562]) (Supplementary Table 3.1). Gene sets tested were Gene Ontology: Biological Processes (GO BP) from *GO.db* Bioconductor package version 3.4.2 (Ashburner et al. 2000). We filtered gene sets to only use those with more than 15 and less than 2000 genes, as small gene sets have very little statistical power and large gene sets tend to be too vague to have meaningful biological interpretation. This resulted in 5159 GO BP gene sets.

Measuring peak distances to nearest transcription start site or enhancer midpoint

Each peak's "regulatory proximity" was defined as the distance, in base pairs, between the peak's midpoint and either the closest TSS or the midpoint of the closest enhancer region. Human gene TSS locations were obtained from the *chipenrich* package, which for hg19 version 3.5.0 are from Bioconductor packages *TxDb.Hsapiens.UCSC.hg19.knowngene* version 3.2.2 (Carlson and Maintainer 2015) and *org.Hs.eg.db* version 3.5.0 (Carlson 2018). Enhancer regions were defined by the union of two sets: DNase hypersensitive sites (DNase DHSs) found in at least two of the 125 cell and tissue types processed by ENCODE (Thurman et al. 2012); and distal and non-promoter DHS within 500kb of the correlated promoter DHSs from 32 cell types (Thurman et al. 2012). The minimum of two cell types was used to reduce false positives. Unions were calculated using the *expand_and_resect2* function in the *granges* R package with min.gapwidth=0, and distal and non-promoter elements were defined as those > 5kb from a TSS. That is, we removed only the portion of an enhancer that was < 5kb from a TSS. This resulted in a total set of 1,616,520 regions > 5kb from a TSS composed of enhancers, silencers, and insulators, although for simplicity we refer to the total set as enhancers. Finally, all peaks are then assigned to the gene with the nearest TSS.

ProxReg step 1: Normalizing for gene locus length and average distance to enhancer

Identical to our previous work, we define a gene's locus length (in bps) as the length of the region on the genome such that a peak binding in the region is assigned to that target gene (Welch et al. 2014); (Cavalcante et al. 2014). As genes with larger locus lengths (i.e. longer distances to neighboring genes) are more likely to have peaks binding farther away from the gene's TSS, gene locus length is associated with average peak distance to TSS, and thus gene locus length is a potential confounding variable. To empirically normalize for gene locus length, we used the combined set of peaks from all 90 ENCODE ChIP-seq peak datasets and computed a cubic

smoothing spline for log locus length (x-axis) vs. log peak distances (y-axis) using the *gam* function in the *mgcv* package. The spline provides the expected, global average binding distance for each gene (D_{spline}), which we then used to obtain the normalized adjusted binding distance (D_{tss}^{adj}) as:

$$D_{tss}^{adj} = \ln D_{tss} - \ln D_{spline}$$

Thus, peaks that are closer to a TSS than expected based on the spline fit will contribute to significant promoter proximity for a gene set.

Similar to how a gene with a longer locus length tends to have peaks farther from its TSS, gene loci with farther spaced enhancers tend to have peaks farther from them. More specifically, the distance to an enhancer region is associated with how far apart a gene's enhancers are spread, which is dependent on both the gene locus length and the number and distribution of enhancers associated with the locus region. Therefore, the average (or expected) enhancer density for each gene is a potentially confounding variable. To normalize for this, we want to first find an average peak distance to an enhancer for every gene. We start with an empirical distribution of peak locations across the genome by combining our list of 90 ENCODE ChIP-seq datasets. We then calculated each peak's distance to the nearest enhancer, assigned all peaks to the gene with the nearest TSS, and finally averaged this distance for each individual gene and use this value as an empirical adjustment for enhancer density ($AvgD_{enh}$). As these 90 experiments do not cover every gene, if a dataset happens to have a peak assigned to a gene not covered, the average distance to enhancer will be set as the predicted mean of a linear estimation using the log gene locus length of the known genes. Similar to the locus length normalization, we have the adjusted enhancer distance (D_{enh}^{adj}):

$$D_{enh}^{adj} = \ln D_{enh} - \ln AvgD_{enh}$$

Thus, peaks closer to an enhancer than expected by chance will contribute to significant enhancer proximity for a gene set.

ProxReg step 2: Testing for proximal regulatory binding

For a gene set of interest, the peaks assigned to genes in the gene set are placed in one group while all other peaks assigned to other genes, called the background genes, are placed in another. We let any gene that has the potential of a peak being assigned to it and annotated in the gene set database to be a background gene, which is equivalent to the procedure of gene expression tools such as DAVID (Huang et al. 2007). The goal is to test whether the peaks in the gene set are significantly closer to TSSs (or enhancers) than expected by chance, given the adjusted distances described above. We use a two-sided Wilcoxon rank-sum test, with positive values denoting the distances in the gene set are *smaller* than those not in the gene set, to test if peaks in the gene set tend to be closer or farther from regulatory regions than those not in the gene set. To account for multiple testing, we use the Benjamini-Hochberg method to calculate FDR values (Benjamini 1995).

Gene set enrichment testing using Poly-Enrich

We tested all 90 ENCODE datasets using the *polyenrich* method in the *chipenrich* Bioconductor package, using the 'nearest_tss' gene locus definition and GO biological processes for the gene sets. Poly-Enrich performs gene set enrichment on sets of peaks by testing if the number of peaks regulating a gene set is greater or less than that not in the gene set, taking into account the number of peaks assigned to each gene (Lee et al. 2018). The statistical model uses a negative binomial *glm* with an adjustment for gene locus length. Significantly enriched gene sets have more peaks, while depleted ones have fewer.

Permutations to assess Type I error rate

To test type I error rate of the proxReg method, we simulated a null set of peak distances (i.e. with no gene sets having significant proximal binding) in three ways: (1) by reassigning every peak to a random gene, where all genes are equally likely to be assigned (*Unif*). (2) to test for correct normalization of gene locus length, we randomized peaks to a gene as above, except genes were first binned with other genes of similar locus length as defined by their TSSs. Specifically, we ranked genes

by locus length, binned them into sets of 100 genes, and then reassigned every peak to a random gene within the same bin (*ByLocusLength*). (3) to test the normalization of average distance to enhancers, we ranked genes by expected distance to enhancer by chance, and then binned genes into sets of 100. Again, we then reassigned every peak to a random gene within the same bin (*ByAvgDEnh*). In all cases, the peaks' original distances to regulatory regions were preserved. We performed ten randomizations per ChIP-seq experiment and chose α-levels of 0.05 and 0.001 to test for a controlled Type I error rate.

Simulations to estimate power

We simulated significant proximal gene sets by starting from a null set of peaks using the *ByLocusLength* permutation strategy. We then added peaks near the TSSs of genes from a gene set, with the choice of a small (471 genes) or a large (1717 genes) gene set. The number of peaks added was equal to 0.01%, 0.05%, or 0.1% of the total number of starting peaks (4,839) in the null set. The distance was chosen from an exponential distribution with mean d_0, and an equal chance for upstream or downstream. We chose values of 100, 500, 1000 for d_0 to simulate scenarios of closer and farther binding. For each scenario, 200 simulated gene sets were ran.

Clustering for TF regulatory patterns

To investigate the regulatory patterns among all 90 ENCODE ChIP-seq data sets, we performed clustering to classify them. We first applied a p-value cut off (<0.001) for both proxReg (promoter and enhancer) results and Poly-Enrich results. We counted the numbers of points (GO BP terms) in each of 4 regions, defined by the ordered pair of signs for the Poly-Enrich and proxReg effect sizes (Figure 3.3), for both promoter and enhancer results in all 90 data sets. Then, a hierarchical clustering heat map was generated based on the log 2 value of counts from each region. The Euclidean distance metric was used with Ward's minimum variance method for clustering. In addition, we also calculated the Pearson correlation between ProxReg promoter results and enhancer results. Since we propose our method as a complementary method for GSE testing, only signed negative log p-values of

significant GO terms (FDR<0.05) from Poly-Enrich were used for correlation calculations.

Test for the ability of ProxReg to reduce false positives from GSE results

To test the ability of our method to reduce false positives from GSE results, we compared the results of ProxReg and Poly-Enrich together to Poly-Enrich alone, using gene sets for each TF that the TF is likely to regulate. Since no gold standard is available for this, we used the GO BP terms that our 35 TFs were assigned to in the human annotation Bioconductor package *org.Hs.eg.db* (Carlson 2018). Motivation for this derives from the fact that transcription factors do not regulate random sets of genes, but rather a well-coordinated set of genes in order to fulfill a cellular biological goal. Indeed, it's been shown that genes in a GO biological process term tend to be regulated by a common TF (Allocco, Kohane, and Butte 2004, Roider et al. 2009, Qian et al. 2005, Ertosun, Hapil, and Osman Nidai 2016). The cellular biological goal is precisely what GO biological process terms aim to describe, as it is defined as "The larger processes, or 'biological programs' accomplished by multiple molecular activities" (Ashburner et al. 2000), which for TFs in DNA binding. Based on these two facts, it follows logically that TFs tend to regulate genes in the biological processes to which they belong. As an example, the NCBI Gene website, an authoritative source for the properties of genes, states in the main summary of E2F family genes that "the E2F family plays a crucial role in the control of cell cycle". This family includes members E2F1, E2F2, E2F3a, E2F3b, E2F4, E2F5, E2F6, E2F7, and E2F8. In each case, we can also find at NCBI Gene that these TFs are assigned to the GO BP terms related to cell cycle. To further validate our approach, we tested whether the TFs actually do tend to target the promoters of genes in their assigned GO terms. Indeed, we found a strong overall trend to targeting more genes in the assigned GO terms versus the non-assigned GO terms (S*upplementary material*). Although TFs may not regulate all of their target gene sets in every cell type, we conclude that the degree of overlap between a method's predictions and a TF's assigned GO BP terms represents a useful benchmarking tool.

To assess this, we first counted all significantly enriched gene sets from Poly-Enrich for all 90 ENCODE ChIP-seq data sets and found their overlap with the GO BP terms each TF was assigned to in *org.Hs.eg.db*. These GO terms were used to count significant GO terms from ProxReg promoter and enhancer results. Fisher's exact test was used to determine whether ProxReg further enriched the resulting GO terms to those assigned to by the TF, beyond what GSE testing accomplished. We used datasets for TFs that are assigned to at least five GO BP terms that were also significant with GSE testing alone. Fisher's exact test results demonstrated whether ProxReg was able to increase the odds ratio of identifying GO BP terms assigned to the TF, compared to GSE testing alone.

Website implementation and Bioconductor availability

ProxReg is available in the *chipenrich* Bioconductor package with the *proxReg()* function, and at the ChIP-Enrich website: http://chip-enrich.med.umich.edu, as an additional option following any of our current gene set enrichment tests. To run ProxReg, the user uploads a file of peaks, which can be in narrowPeak or BED format. They then select to test for proximity to either nearest TSS (NTSS) or enhancers. Currently, we have only implemented testing for enhancer proximity in human (hg19 genome), but others will be added as enhancers are sufficiently defined in other species and newer genome versions. Finally, the user selects what gene sets to test from any of our included gene sets (details in *chipenrich* package and on website) or a user-generated set. An example of

The *proxReg()* function outputs four files:

opts: The options that the user input into the function.

peaks: A peak-level summary showing the peak-to-gene assignment for each peak, as well as their distances to TSS or enhancer.

results: The results of the proximity tests. Lists the tested gene sets along with their descriptions, the test effect, closer/farther status, p-value, and FDR. Also included is the list of Entrez gene IDs with contributing signal for each proximity test.

qcplot: A histogram showing the distribution of peak distances.

All R code for recreating analysis and figures can be found at: https://github.com/sartorlab/proxReg. An example for the use of ProxReg can be found in the *chipenrich* Bioconductor vignette.

Results

Overview of ProxReg method

We developed a new method, ProxReg, to test the proximity of peaks to transcription start sites (TSSs) or enhancers in a gene set of interest. The motivation for our new method is illustrated in Figure 3.1. The goal is to test whether the enrichment of a GO term or pathway is driven by regulation via promoters or distal regions (i.e. enhancers). To accomplish this, we firstly measure the distances from the midpoints of the peaks to the nearest regulatory regions (either TSSs or enhancers), and assign each peak to its target gene according to the gene with the nearest TSS (Welch et al. 2014). Specifically, for each gene we defined its gene locus to be the region between the upstream and downstream midpoints of its TSS and the neighboring gene's TSSs. However, one cannot simply directly test whether the distances are smaller within a gene set versus other genes, due to potentially confounding variables that first need to be taken into account. Since a gene locus with a large length was observed to have farther peaks from its TSS on average (Figure 1A), we first normalize for the gene locus length before testing the proximity to TSSs (see Methods). For enhancers, we observed that the distance to an enhancer was dependent on the average distance from each enhancer to peaks in a gene locus (Figure 1B). Thus, we normalized the raw peak to enhancer distances using the average enhancer density for each gene. Finally, a two-sided Wilcoxon rank sum test was used for testing the proximity of peaks in a gene set to TSSs (or enhancers) compared to peaks outside the gene set. Generally, this test would be performed on all of the enriched gene sets identified by a gene set enrichment (GSE) test, to understand whether the enrichment of each gene set was due to regulatory activity near promoters or enhancers.

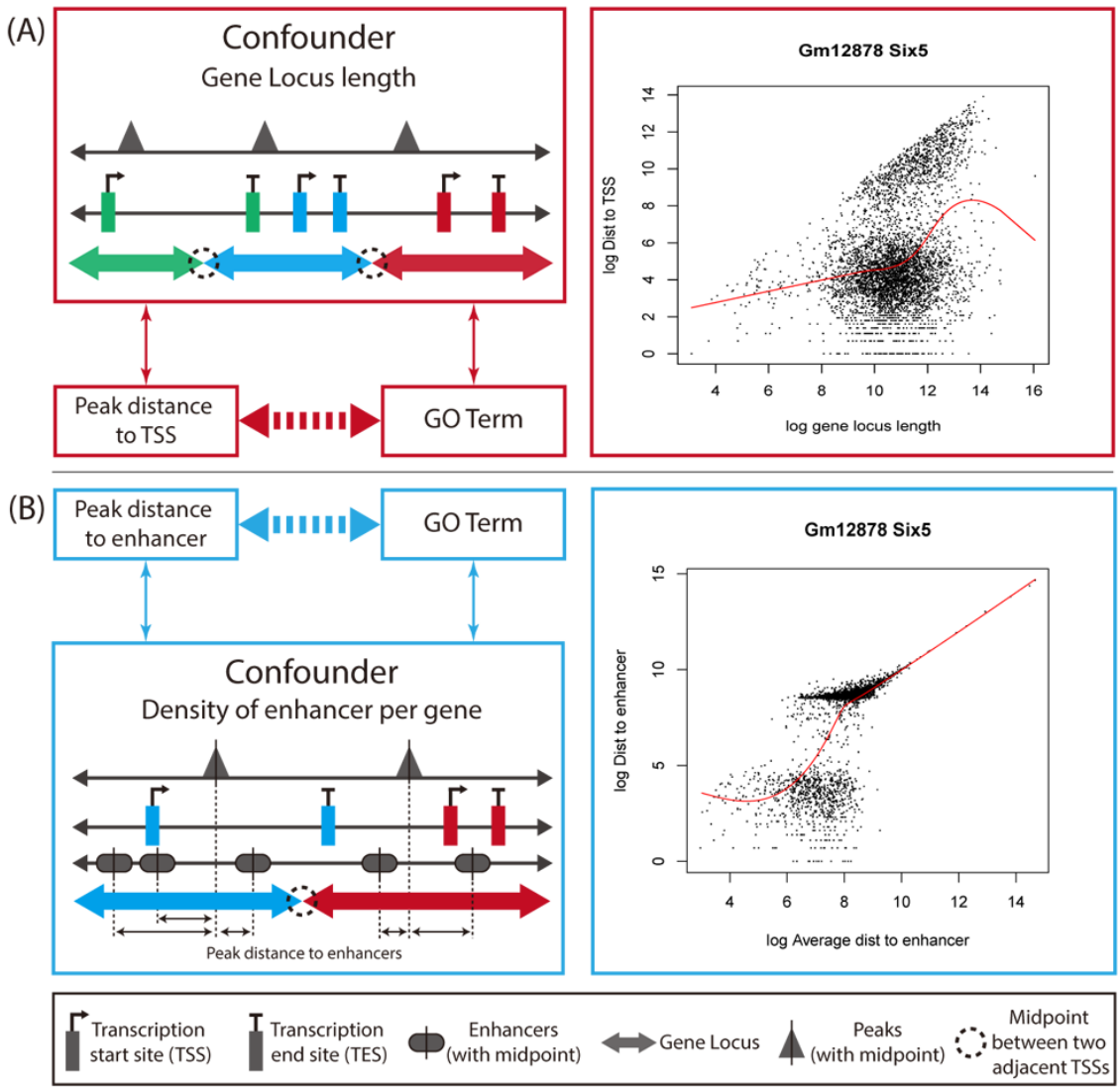


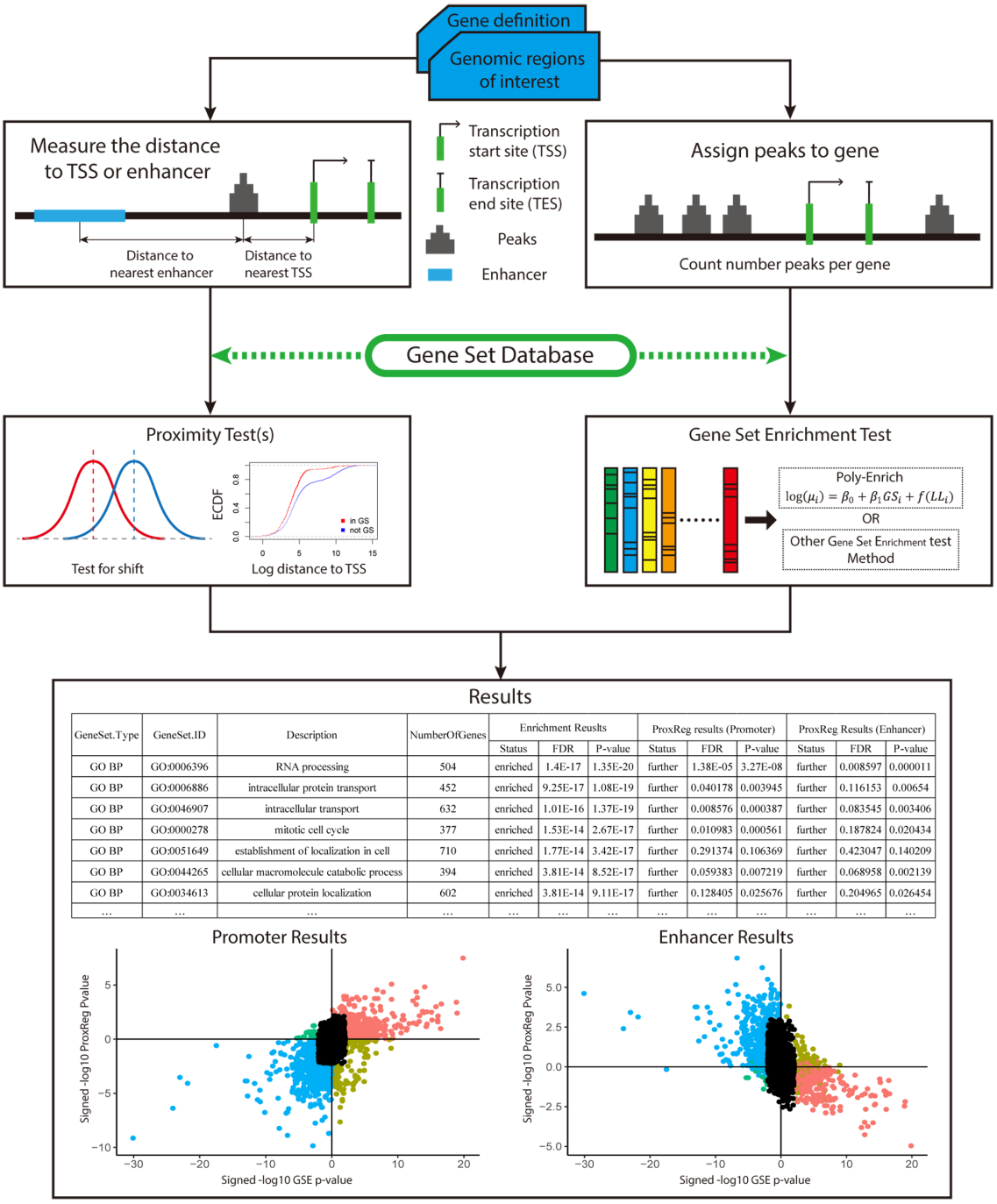
Figure 3.1 Overview of how ProxReg adjusts for confounding variables.

We describe the ProxReg adjustments in two parts. (A) When testing proximity to TSSs, we normalize the peak distances to TSSs according to their relationship with gene locus lengths. (B) When testing proximity to enhancer, we normalize the peak distances to enhancers according to their relationship with enhancer density, modeled by the average distance of any peak to an enhancer. In both cases, we avoid a potential confounding effect, as shown by the arrows between variables on the left-hand side.

Recommended workflow for ProxReg

To test our new method, 90 ENCODE ChIP-seq data sets (36 transcription factor in three Tier 1 cell lines) (Qu and Fang 2013, Consortium 2004, Sloan et al. 2016) were used in this study. The recommended workflow for implementing our new method

is summarized in Figure 3.2. We begin with a gene definition file containing gene locus definitions (provided by our software, or uploaded custom by the user) and a set of peaks of interest (provided by the user). The distance between the midpoint of peaks and nearest TSSs (or midpoint of enhancers) are measured and adjusted for all background genes. The ProxReg non-parametric test is ran for the chosen gene sets (e.g. GO). In parallel to this, a standard gene set enrichment test is performed using the same gene sets. In this paper, we applied the *polyenrich* method for the GSE test (Lee et al. 2018), but others may be used. Result files contain the proximity results with test direction (enriched/depleted from GSE, and closer/farther from ProxReg), p-values and FDR values. Combined with the p-values from GSE, the gene set proximity and enrichment patterns can be easily visualized (Figure 3.2 Results Section).



GeneSet.Type	GeneSet.ID	Description	NumberOfGenes	Enrichment Reuslts			ProxReg results (Promoter)			ProxReg Results (Enhancer)		
				Status	FDR	P-value	Status	FDR	P-value	Status	FDR	P-value
GO BP	GO:0006396	RNA processing	504	enriched	1.4E-17	1.35E-20	further	1.38E-05	3.27E-08	further	0.008597	0.000011
GO BP	GO:0006886	intracellular protein transport	452	enriched	9.25E-17	1.08E-19	further	0.040178	0.003945	further	0.116153	0.00654
GO BP	GO:0046907	intracellular transport	632	enriched	1.01E-16	1.37E-19	further	0.008576	0.000387	further	0.083545	0.003406
GO BP	GO:0000278	mitotic cell cycle	377	enriched	1.53E-14	2.67E-17	further	0.010983	0.000561	further	0.187824	0.020434
GO BP	GO:0051649	establishment of localization in cell	710	enriched	1.77E-14	3.42E-17	further	0.291374	0.106369	further	0.423047	0.140209
GO BP	GO:0044265	cellular macromolecule catabolic process	394	enriched	3.81E-14	8.52E-17	further	0.059383	0.007219	further	0.068958	0.002139
GO BP	GO:0034613	cellular protein localization	602	enriched	3.81E-14	9.11E-17	further	0.128405	0.025676	further	0.204965	0.026454
...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 3.2 Overview of how of ProxReg fits in with the overall workflow of gene set enrichment testing with genomic regions.

The peak distances to TSSs or enhancers are calculated for the proximity test. In parallel, all peaks are assigned to genes for gene set enrichment testing. The same gene set database is used for both proximity and gene set enrichment testing. Combining the gene set enrichment and proximity tests, the results can be visualized as shown in the results section. The left scatter plot is an example of the

combination of enrichment and promoter results. The right scatter plot is an example of enhancer results combined with enrichment results. The X-axis of these two scatter plots represent the gene set enrichment test result. A larger signed -log p-value indicates more enrichment, while negative values indicate depletion. The Y-axis represents the proximity results. Larger signed -log p-values indicate GO terms having genomic regions closer to the TSSs or enhancers.

Controlled Type 1 error rate and ability to detect true positive results

We validated the Type 1 error rate (rate of false positives) of ProxReg using randomizations of real datasets to simulate null datasets with no significant proximities to TSSs or enhancers. We performed three types of permutations: the "*Unif*" permutation, which takes every peak and reassigns another gene to it with each gene having the same probability, the "*ByLocusLength*" permutation, which tests the effectiveness of the locus length normalization in the distance to TSS test, and the "*ByAvgDEnh*" permutation, which tests the effectiveness of the normalization to average distance to enhancer in the distance to enhancer test (see Methods for details). For a p-value < 0.05 cutoff, we expect a Type I error rate of approximately 5%. For a p-value < 0.001 cutoff, we expect a Type I error rate of approximately 0.1%. Results indicate that for each permutation (*Unif* and *ByLocusLength* for TSS proximity tests, and *Unif* and *ByAvgDEnh* for enhancer proximity tests), the Type 1 error rate is reasonably controlled at the expected level (Supplementary Figure 3.1).

To ensure that our method is able to identify gene sets with true cases of TSS or enhancer proximity, we generated artificial peak datasets starting with a randomized data set using the *ByLocusLength* permutation, and then adding peaks with TSS distances following a specified distribution. We added peaks by varying the number of peaks and the distance of peaks to assess a wide range of scenarios. We also used two gene sets of different sizes (see Methods for details). We expected the following changes in parameters to increase power: a smaller gene set used (easier to influence average distance), more peaks added, and a smaller average distance. We can see that all three of these scenarios increased power to detect the true positive gene sets as expected (Supplementary Figure 3.2).

Integration of GSE and ProxReg results reveals different regulatory patterns of TFs

We clustered the 90 ENCODE ChIP-seq datasets into three groups based on the hierarchical clustering heat map illustrated in Figure 3.3. The first and largest group (47 datasets) is characterized by a strong positive correlation between GSE and promoter (TSS) ProxReg signed significance levels, and a strong negative correlation between GSE and enhancer ProxReg signed significance levels, indicating that the majority of enriched gene sets are due to binding near TSSs (Figure 3 blue cluster; many genes in regions p1, p2, e3 and e4). Transcription factors like SIX5 (SIX homeobox 5), SP1 (Specificity Protein 1*), and GABP (Nuclear Respiratory Factor 2) are included in this group. The second largest group (32 datasets) is more interesting because the datasets consist of some enriched gene sets with significant proximity to promoters, and other enriched gene sets with significant proximity to enhancers (Figure 3 red cluster; genes spread out across mainly p1, p4, e1, and e4). The results for these transcription factors enable understanding the different regulatory mechanisms used for different biological processes. MEF2A (Myocyte-specific enhancer factor 2A) in K562 cells, a member of this group, was observed to regulate GTPase activity and translational initiation-related GO terms from TSSs, and transmission of nerve impulse and multicellular organismal signaling GO terms from enhancers. Similarly, P300 (Histone acetyltransferase p300), a well-known marker of enhancers, was found to regulate chromatin organization from TSSs, while regulating phosphatidylinositol dephosphorylation and phosphatidylinositol-mediated signaling-related GO terms from enhancers (Fryer et al. 2002, De Luca et al. 2003). The smallest group included only 11 ChIP-seq datasets. This group was characterized mainly by enriched gene sets with many having significant proximity to enhancer regions and/or far from promoters (Figure 3.3 purple cluster; many genes in p4 and e1). Members of this group included Pol II in all three cell lines and EGR1 in K562 cells and Gm12878 cells, indicative of Pol II binding along entire gene lengths and not just at promoters. In addition, we examined the Pearson correlation between promoter results and enhancer results for all 90 ENCODE ChIP-seq data

sets. Eighty-eight of them show a negative correlation between the promoter results and enhancer results (Figure 3.4). This negative correlation indicates that overall, GO terms are significantly enriched either by the TF binding closer to promoters or closer to enhancers. Among these 90 data sets, most of them (67 out of 88 data sets) show a strong negative correlation as shown in Figure 4A. Several of them have weak correlations as shown in Figure 4B. The two datasets that did not show negative correlations are NRSF and CMYC in H1-hESC cells. After removing non-significant GO BP terms from Poly-Enrich results, NRSF data set shows a weak positive correlation based on the remaining GO BP terms and no significant GO BP terms in CMYC data set.

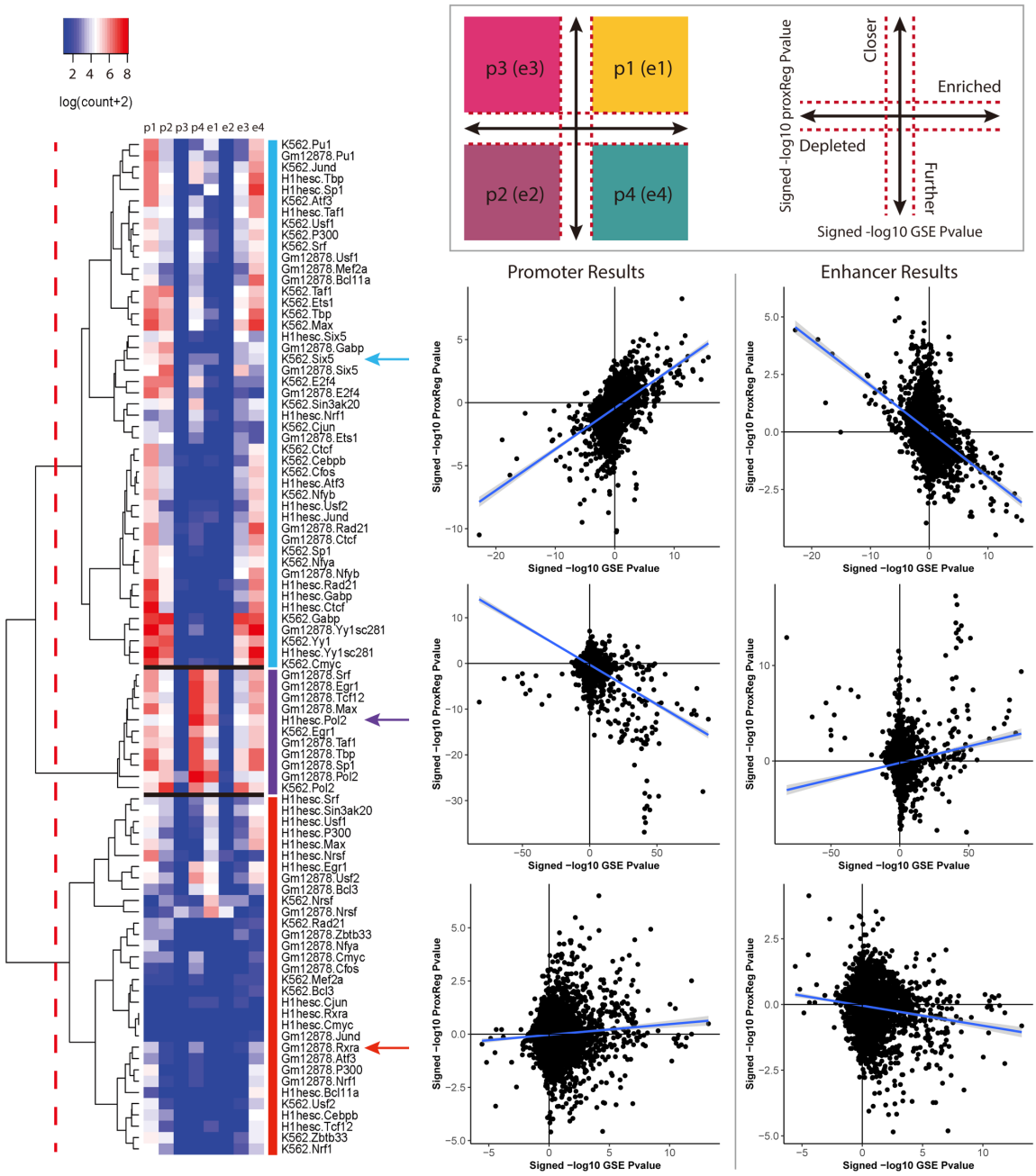


Figure 3.3 The regulation patterns of the 90 ENCODE ChIP-seq datasets.

A p-value cutoff (<0.001) was applied to define the four regions as illustrated in the top panel. The cutoffs are represented by the red dash lines. For each data set, the points count of the combination of ProxReg promoter results and Poly-Enrich results are labeled as p1, p2, p3 and p4. Similarly, the combination of enhancer results and Poly-Enrich results are labeled as e1, e2, e3, and e4. Based on our analyses, 47 data sets show a clear positive correlation in promoter results and a clear negative correlation in enhancer results. 32 datasets show no strong

correlation in either promoter or enhancer results. The remaining 11 data sets show a clear positive correlation in the enhancer results. For each group, the promoter and enhancer results of one data set are illustrated as an example.

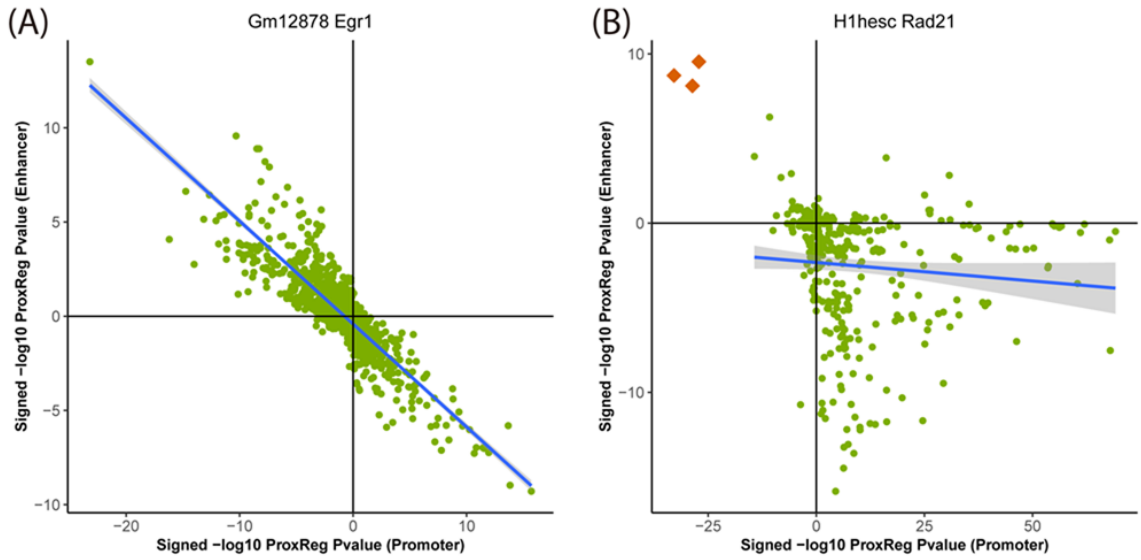


Figure 3.4 Examples of the correlation between ProxReg promoter p-values and enhancer p-values.

Majority of the 90 ENCODE ChIP-seq data sets show a strong negative correlation as shown in (A). A small portion of these data sets show a pattern as shown in (B). The three orange dots in (B) are GO terms related detection of chemical stimulus (GO:0050907, GO:0009593, and GO:0050911).

ProxReg identifies known associations with promoter and enhancer binding, using SIX5 and NRSF peaks

To further illustrate our method, we assess ProxReg results for two TFs known to have a very strong tendency to bind either in proximal promoters or enhancers. We first selected SIX homeobox 5 (SIX5) in GM12878 cells as an example, which is involved in determination and maintenance of retina formation that proposed binding to promoter regions of related genes (e.g. myogenin and IGFBP5) (Sato et al. 2002, Spitz et al. 1998). The results of SIX5 are shown in Figure 3.5.

In Figure 3.5A, we can see that the majority of the ChIP-seq peaks (67.4%) are near TSSs. Through the combination of ProxReg results and Poly-Enrich results, a great majority of gene sets are enriched by the transcription factor binding near TSSs

(positive correlation in Figure 3.5B) instead of near enhancers (negative correlation in Figure 3.5C). Using two particular GO terms from the scatter plots, we show the distribution of distances from peaks to TSSs or enhancers (bottom part of Figure 3.5B and 3.4C). Combining the locations of these two GO terms (GS1 and GS2 in the scatter plots), illustrates how our method is able to provide additional information for interpreting GSE testing results.

We also selected Neuron Restrictive Silencer Factor (NRSF) in the K562 cells as an example. NRSF, also known as RE1-Silencing Transcription factor (REST), is a transcription factor known to silence neuronal genes in non-neuronal cells, it can act as a transcriptional repressor or enhancer of target genes, often regulating from enhancer regions (Schoenherr and Anderson 1995, Seth and Majzoub 2001). Almost half of NRSF ChIP-seq peaks (51.2%) are far from TSSs (Figure 3.5D). A similar strategy was used for illustration of ProxReg with the transcriptional repressor NRSF in K562 cells. Consistent with previous observations that this transcription factor tends to bind to silencers/enhancers instead of promoters, there is a relative strong positive correlation shown in the enhancer scatter plot (Figure 3.5F) but not for TSSs. Thus the results confirm that most enriched GO terms were enriched due to the transcription factor binding in or near enhancer regions. These results validate our new method, ProxReg, is a powerful tool that can be used as a complementary approach for interpreting GSE test results.

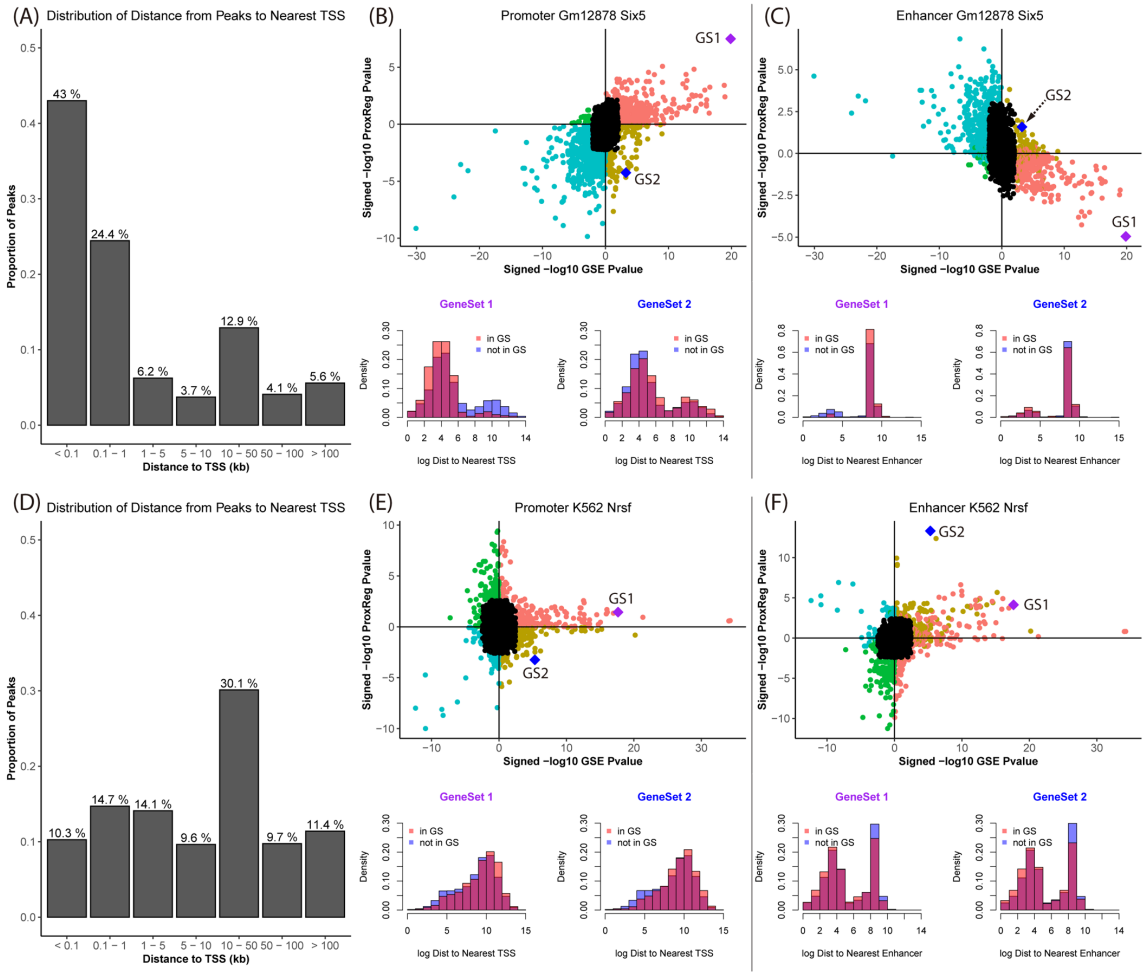


Figure 3.5 Illustration of ProxReg results.

The results of SIX5 in GM12878 cell lines are shown in Figure 3A-C. (A) the distribution of distances from peaks to nearest TSSs. (B) scatter plot of the combination of enrichment results and promoter results. Two gene sets were selected to show the distance distribution to nearest TSSs for genes in the gene set and not in the gene set. (C) enhancer results combined with the enrichment results. The same gene sets were used in this scatter plot. The distribution of distances to the nearest enhancers of these two gene sets are shown in the bottom of Figure 3C. Similar to SIX5 results, Figure 3D-F show the results of NRSF in K562 cells. For SIX5, GeneSet 1: RNA processing. GeneSet 2: Positive regulation of nitrogen compound metabolic process. For, NRSF, GeneSet 1: Neuron differentiation. GeneSet 2: System process.

ProxReg enriches GSE findings for likely true positives

We assessed whether ProxReg can be used to not only estimate the proximity effects but also help users to remove possible misleading or false positive gene sets from GSE results. To accomplish this, we compared the significantly enriched gene sets to a set of GO biological process (BP) terms from *org.Hs.eg.db* for each TF before versus after taking into account their ProxReg results. The GO BP terms from *org.Hs.eg.db* consists of the TFs and the assigned GO BP terms for the gene that encodes them (see Methods for more detail).

We used ChIP-seq datasets with at least five significantly enriched GO BP terms in their *org.Hs.eg.db* set (to ensure sufficient power), which resulted in 28 datasets with ProxReg enhancer results and 36 datasets with ProxReg promoter results. We then tested whether requiring a significant ProxReg test resulted in a higher odds ratio of detecting the *TF-assigned* GO BP terms. Of the 28 enhancer dataset results, 18 (64%) had an odds ratio greater than 1. Among these, 11 (61%) of them were significant. Conversely, only 2 enhancers' results had an odds ratio significantly less than 1. These two results were from EGR1 and ATF3 in K562 cells. Previous research (Cullen, Brazil, and O'Connor 2010) suggests that EGR1 recognizes and binds to promoter regions of target genes, so it is possible that the GO BP terms from *org.Hs.eg.db* we compared to is incomplete, with previous data mainly being focused on biological processes that EGR1 regulates from promoter regions. A similar case may be true for ATF3.

Among 36 ProxReg promoter results, 25 (69%) had an odds ratio greater than 1. Among these 25 results, 15 (60%) of them were significant. Conversely, only 2 promoter results were significant with an odds ratio smaller than 1. One of them was PU.1 in K562 cells. A previous study (Heinz et al. 2015) indicated that PU.1 usually binds to a PU-box found on enhancers of target genes, consistent with the ProxReg promoter results of PU.1 peaks having an odds ratio less than one. Although we only found 5 significant GO BP terms from our results that are also assigned to PU.1, some other significant GO BP terms that we identified were biologically related to the remaining GO terms assigned to PU.1. For instance, some

GO terms assigned to PU.1 were related to response to toxic substances, drugs, and antibiotics, and many immune response-related GO terms were significant. Overall, these results demonstrate that ProxReg can be used as a powerful supplemental method to remove misleading or false positive GSE test results (Supplementary Table 3.2), and provide additional evidence for novel regulated processes initially identified by GSE testing.

ProxReg analysis identified NRSF regulatory pattern switching in different cell types

The ProxReg results can guide and refine the biological interpretation of GSE results by identifying whether each enriched gene set is regulated mainly via binding close to promoters or enhancers. We exemplified this using the findings of NRSF, which was shown to regulate neuron development mostly via binding to enhancers in K562 cells (see details above). To further investigate the regulation patterns of NRSF in different cell lines, we utilized ENCODE NRSF ChIP-seq experiments from three cell types (GM12878, H1-hESC and K562), and performed and integrated the Poly-Enrich and ProxReg analyses for each cell type. In GM12878, almost all significant GO terms identified by both Poly-Enrich and ProxReg were found to be closer to enhancers, except one GO term "establishment of localization in cell", which was significantly closer to promoters (FDR = 2.04×10^{-6}) and farther from enhancers (FDR = 9.60×10^{-7}) (Figure 3.6A and 3.6B, Supplementary Table 3.3). Most of them were related to neuron development, including "neurological system process", "regulation of nervous system development", and "synapse organization". In H1-hESC cells, however, NRSF binding sites were significantly enriched in GO terms which were significantly closer to promoters, and mostly related to neuron development and regulation, such as "synapse organization", "neuron projection guidance" and "neurotransmitter secretion" (Figure 3.6A and 3.6C, Supplementary Table 3.4). Less than 1% GO terms were closer to enhancers ("cell morphogenesis involved in differentiation", "regulation of cell projection organization", and "positive regulation of nervous system development"). The pattern observed in K562 was similar to that in GM12878: the majority of enriched GO terms were

significantly closer to enhancers, and again most of them were related to neuron regulation (e.g. “axon guidance”, “synapse maturation”, and “regulation of synapse assembly”) (Figure 3.6A and 3.6D, Supplementary Table 3.5), whereas only one was closer to promoters (“regulation of alternative mRNA splicing, via spliceosome”). These findings point to a fundamental shift in the binding patterns of NRSF to regulate neuronal genes during neuron development and organization processes: closer to promoters of genes in embryonic stem cells (H1-hESC), while closer to enhancers in differentiated cells (GM12878 and K562). Taken together, we demonstrate that ProxReg analysis complements the GSE results by distinguishing where a TF binds to regulate genes, which is key to understanding the mechanisms of gene regulation and guiding potential targeted gene therapy. ProxReg is incorporated in the *chipenrich* Bioconductor package and ChIP-Enrich website, and can be used with many additional databases of gene sets.

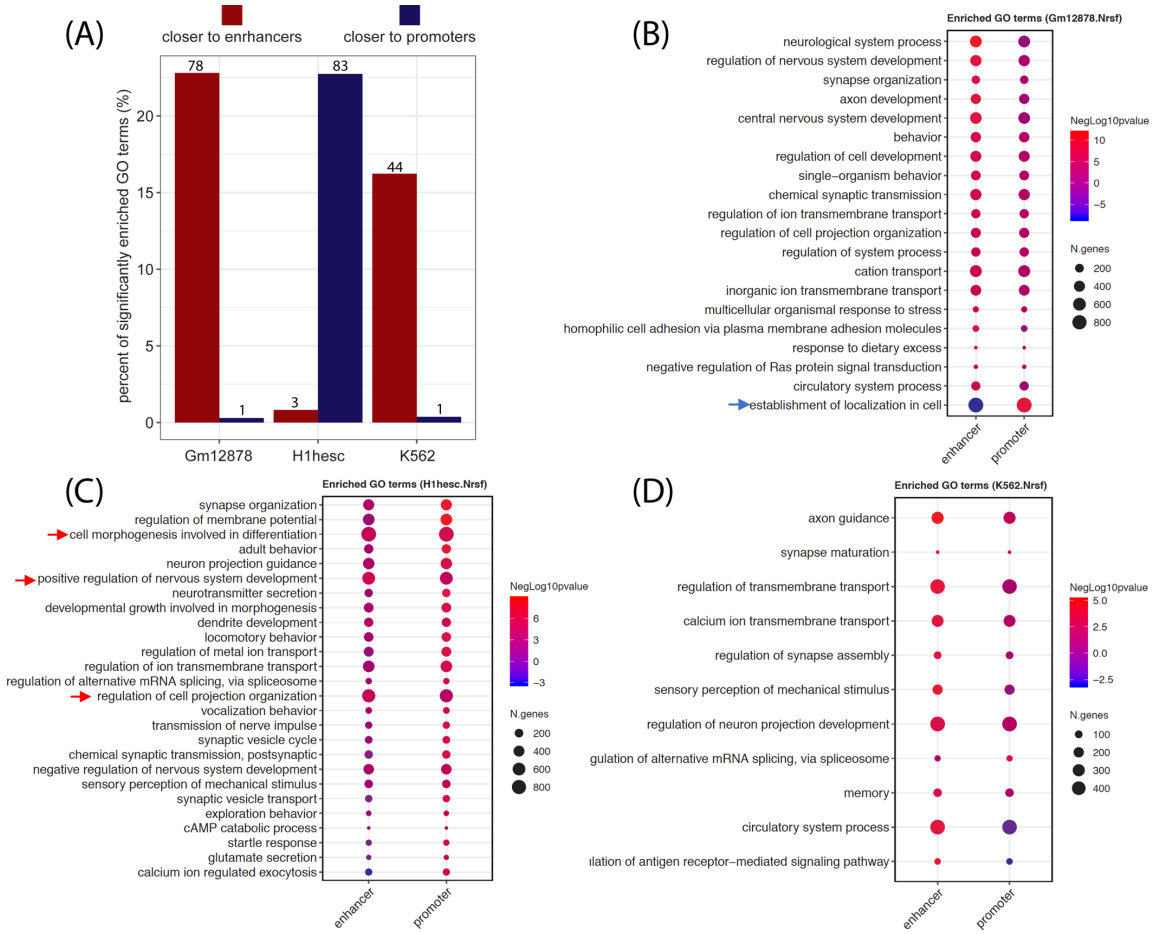


Figure 3.6 The different regulatory patterns of NRSF in three cell lines.

(A) The bar plots show the percentage of significantly enriched GO terms that were closer to enhancer (dark red) or promoter (dark blue) in each cell line (x axis). The numbers of terms were marked on the top of each bar. (B)-(D) The dots represent the ProxReg enhancer or promoter significance levels (signed negative log p-values, resulting in positive values for proximal regions, and negative values for more distal regions) of the enriched GO terms in GM12878 (A), H1-hESC (B) and K562 (C) cell lines. In a particular cell line, the arrows point to the GO terms closer to promoters (blue arrows) while most of the terms are closer to enhancers, or point to the GO terms closer to enhancers (red arrows) while most of the terms are closer to promoters. For visualization, the redundant GO terms were removed from the list (Koneva et al. 2018).

Discussion

We introduced a genomic region proximity test method called ProxReg that can be used as a complement for gene set enrichment (GSE) tests. The standard GSE tests for sets of genomic regions (e.g. ChIP-seq peak sets) usually only consider the relationship between the genomic regions and TSSs (McLean et al. 2010). However, it is of great interest to know whether a gene set is significantly enriched through regulatory activity near promoters or enhancers. Our new method, ProxReg, is able to find gene sets with regions that bind significantly closer to (or farther from) either promoters or enhancers. Furthermore, we validated that it has an appropriate Type I error rate, and that the statistical power of the test behaves as expected when varying the relevant variables. ProxReg uses a two-sided Wilcoxon rank-sum test for the proximity test while adjusting for important confounding variables. On its own, it provides insight into particular regulatory patterns. Integrated with GSE testing, it serves as a powerful complementary approach to enhance understanding of regulatory behavior across cell types, time points, disease stages, and more.

When performing pathway analyses with current tools, the method may detect significance from regulation coming from different regions, but the underlying details are often left unknown. Standard GSE tests either do not take proximity to regulatory regions into account, or embed the proximity to TSSs within the test, still ignoring enhancers. In this way, it is difficult to interpret the results without the

proximity effects. For example, when GREAT or Poly-Enrich finds a significant gene set from a ChIP-seq experiment, it is known that the gene set is enriched with peaks compared to genes not in the gene set, but we do not know if the peaks reside in promoter or enhancer regions any more than expected by chance. ProxReg is able to further show if the binding sites are closer to (or farther from) TSSs or enhancers, giving more insight into a transcription factor's binding tendencies. We showed with real world ChIP-seq datasets from ENCODE that ProxReg was able to identify tendencies of transcription factors known to most often bind in proximal promoter regions (SIX5) (Spitz et al. 1998, Sato et al. 2002) or distal regions (NRSF) (Schoenherr and Anderson 1995, Seth and Majzoub 2001). Additionally, significantly enriched gene sets that were not found to be significant by ProxReg may have resulted from distal peaks being misassigned to incorrect target genes. To illustrate the usefulness of ProxReg, we performed GSE and ProxReg testing on three ChIP-seq datasets of the transcription factor NRSF in embryonic stem cells (hESC) and two differentiated cell lines (K562 and GM12878). We showed how NRSF tends to regulate certain neuronal-related gene sets in differentiated cells by binding closer to enhancer regions, while regulating similar gene sets via binding to promoters in embryonic stem cells. Furthermore, we identified other non-neuronal GO terms that NRSF regulates via binding mainly in promoter (or enhancer) regions. It is interesting to note that the enhancer binding, which is more cell-type specific and generally evolved later than regulation from promoters (Cai et al. 2019, Nord et al. 2013), was identified for the complex neuron development and related terms, while more basic processes such an establishment of location in cell and mRNA splicing, were regulated from closer to TSSs. Only in embryonic stem cells was even the neuronal-related terms regulated via promoters.

A particular point of interest is the peaks that do not bind close to either TSSs' or enhancers. One possible explanation is that those peaks are miscalled false positives and there is actually no binding activity in those regions. However, it is also possible that there are undiscovered enhancers, which should classify those peaks as close to an enhancer, but are instead evaluated as farther. Additionally, if the true biological

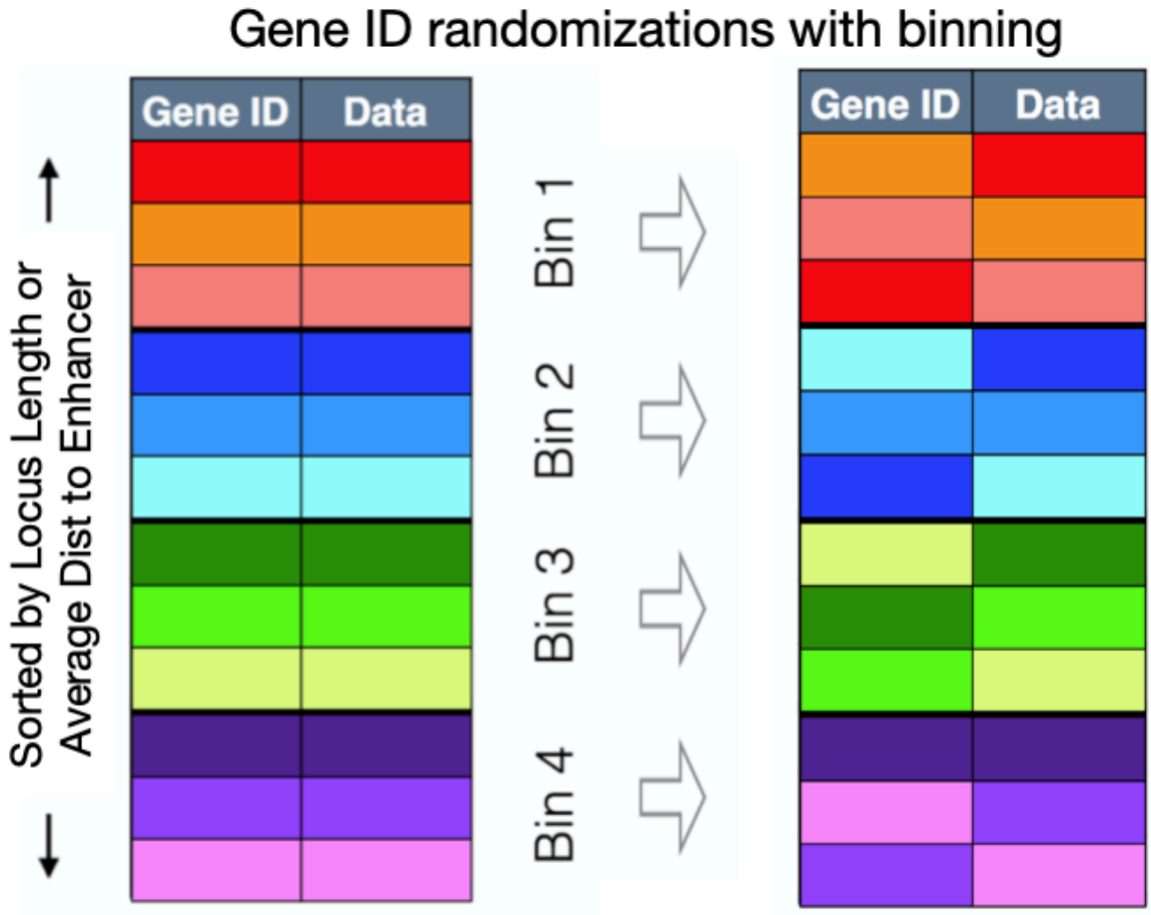
mechanism of distal regulation requires a certain proximity to an enhancer region, then it may be of interest to only consider peaks within a certain base pair threshold of enhancer midpoints.

ProxReg has multiple limitations. Currently, we have implemented distance to enhancers for human (hg19 genome), and have provided support for hg38 (Haeussler et al. 2019) using LiftOver (Kuhn, Haussler, and Kent 2013). Since the enhancer landscape for other organisms lags the comprehensiveness of that for humans, we currently only offer the promoter proximity test for other species. As other organisms' enhancer locations become more accurately defined, we plan to add support for more enhancer proximity tests.

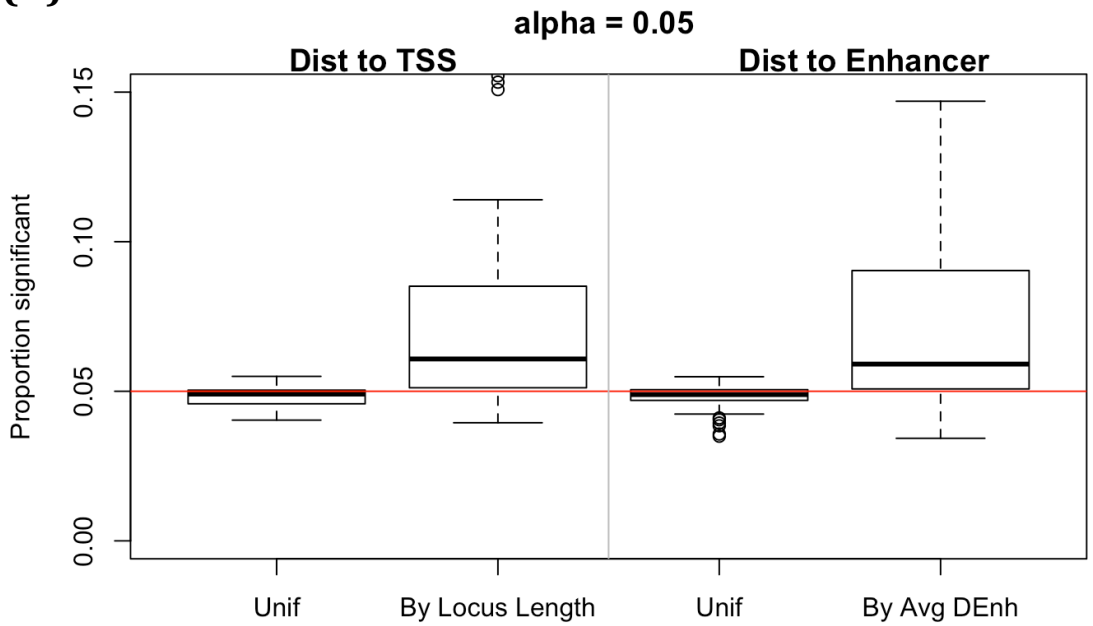
An ongoing question is the identity of the targeted genes of enhancers binding events (Melamed et al. 2016, Sanyal et al. 2012, Rubtsov et al. 2006), which remains challenging due to long-range chromosome interactions. By analyzing transcription factors that tend to bind far from TSSs, we found that there are gene sets that tend to be regulated by TFs binding significantly farther from gene TSSs while also binding closer to enhancer locations. However, ProxReg assumes that each peak is associated with the gene with the nearest TSS, whereas this is often not true. It has been estimated that 79-95% of transcription factor binding actually regulates a gene interceded by one or more other genes (van Heyningen and Bickmore 2013, de Sotero-Caio et al. 2017, Aldrup-Macdonald and Sullivan 2014). Additionally, we used one general set of enhancer locations across the entire genome, whereas in reality, this method may benefit from allowing different tissues to have different sets of defined enhancer locations. Further research is required to understand how the comprehensiveness of the enhancer database affects the results of ProxReg, as well as of GSE tests. We are currently undergoing research on the differences in enhancer locations and their target genes in relation to GSE testing.

Supplementary Figures for Chapter 3

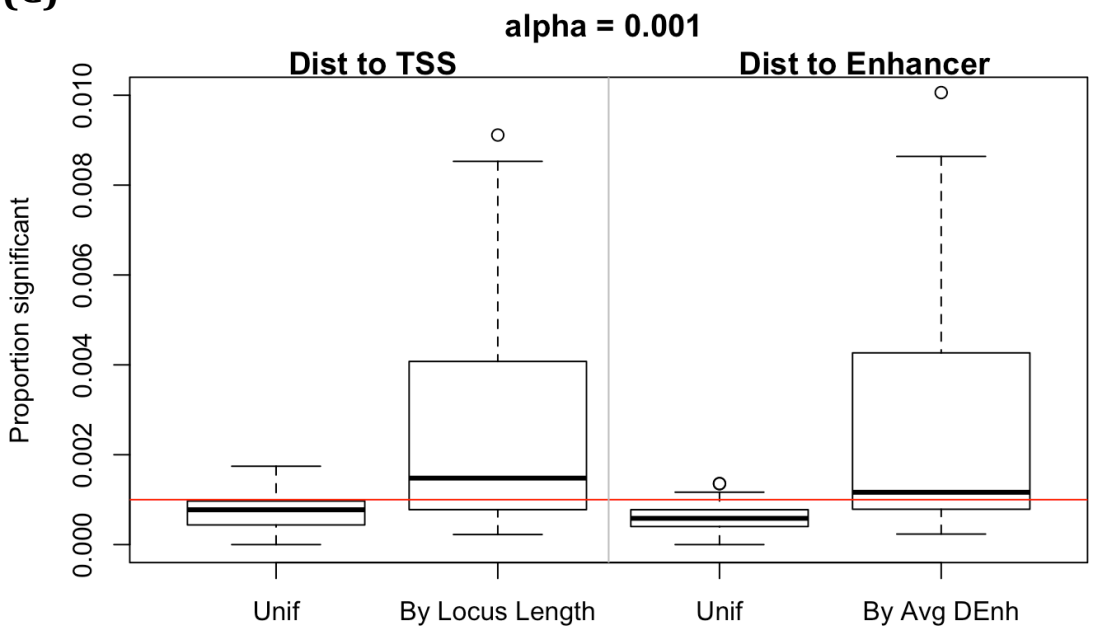
(A)



(B)

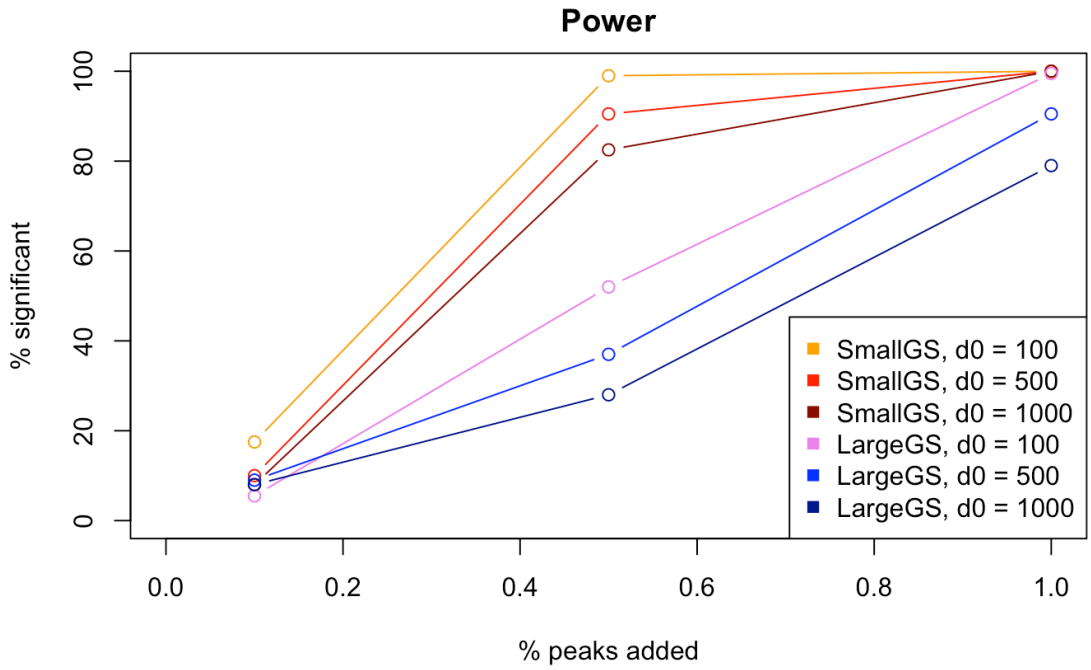


(C)

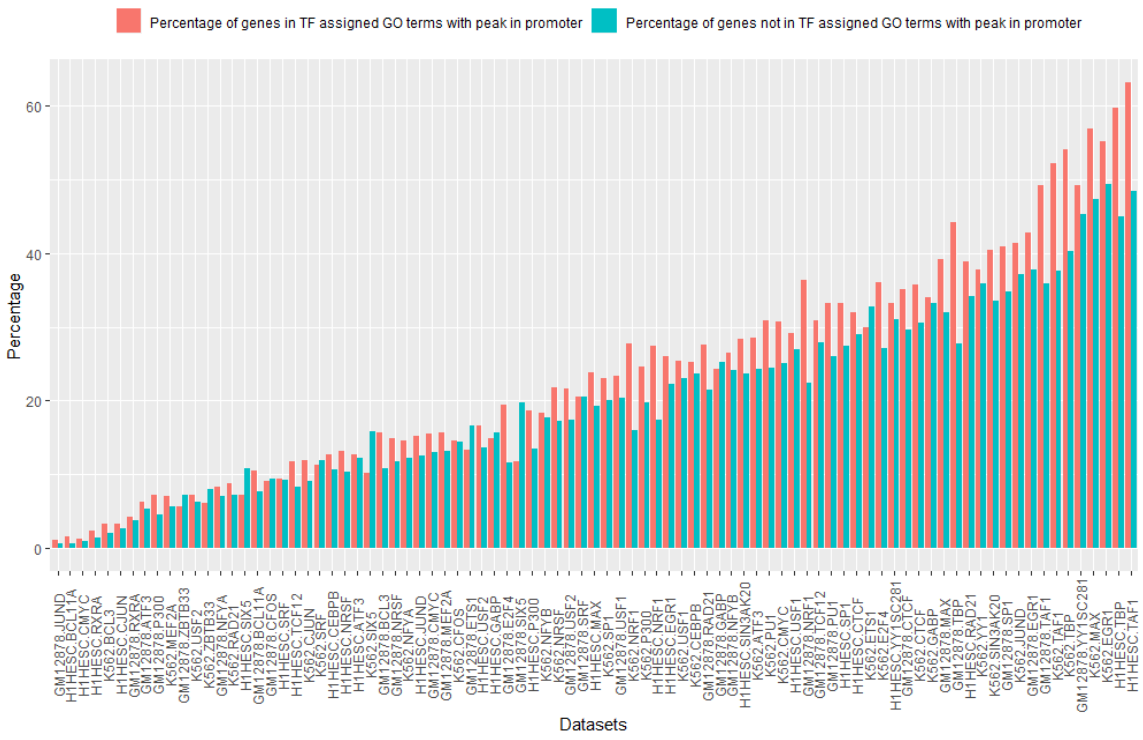


Supplementary Figure 3.1.
Type I error simulation results. (A) We permuted the peak-to-gene assignments in three ways: Unif is reassigning each peak to another gene with every gene having the same chance; By Locus Length first bins all the genes into bins of similar locus length, then the peak is assigned a gene uniformly from the bin of the gene it was originally assigned to; and By Avg DEnh is similar except the genes are binned by empirical average distance to enhancer. (B and C) Each point in the box plots is the

proportion of permuted gene sets under a specific threshold (0.05 for B and 0.001 for C) for each of 90 transcription factors. The Unif randomizations are well controlled, and there are some outliers for the By Locus Length and By Avg DEnh randomizations, but the median Type 1 error is still relatively well controlled.



Supplementary Figure 3.2.
Power simulation results. We generated pseudo-enrichment data by first starting with a permuted peak set in bins of locus length, and then added peaks to genes to a particular gene set. We chose a small (471 genes) and a large (1717 genes) gene set. The number of peaks added were 0.01%, 0.05%, or 0.1% of the total number of peaks in the experiment, which was 4839. Peak distances were added based on the following distribution: $P(Dist = x) = \exp\left(-\frac{x}{d0}\right)$ with d0 being the average distance from the TSS and the choices of d0 being 100,500, or 1000. As expected, when more peaks were added, peaks were added closer, or the smaller gene set was used (higher proportion of closer peaks in the gene set), the power increased.



Supplementary Figure 3.3.
To support our assumption that TFs tend to regulate genes in the biological processes to which they belong, we compared the percentage of genes with a ChIP-Seq peak in their promoter (a) in the assigned GO BP terms, with (b) not in the assigned GO BP terms for each TF. Only promoter regions were used to be confident in the peak-to-gene assignments. The GO BP term assignments for each TF were extracted from the human annotation Bioconductor package *org.Hs.eg.db.* Most of the ChIP-seq data sets used in this paper show a higher percentage of genes in TF-assigned GO terms have a peak in their promoter than genes in other GO terms. GO terms with less than 1000 genes were used in this figure. We also performed a proportion test for the difference for each TF, and the results can be found in Supplementary table 3.6.

Supplementary Tables for Chapter 3

Supplementary tables can be found in my Github dissertation repository at https://github.com/leetaiyi/Dissertation

Supplementary Table 3.1: List of all 90 ENCODE datasets used for gene set enrichment testing. Downloaded from ENCODE Analysis data at UCSC: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/

Supplementary Table 3.2: Using proxReg to identify false positive Poly-Enrich test results by seeing how many true positives were found

Supplementary Table 3.3: A merged proxReg and Poly-Enrich results file for Gm12878 NRSF. Most results where both methods are significant were for proximity to enhancers.

Supplementary Table 3.4: A merged proxReg and Poly-Enrich results file for H1-hesc NRSF. In contrast to Gm12878 shown in Supplementary Figure 3.3, most results where both methods are significant were for proximity to promoters.

Supplementary Table 3.5: A merged proxReg and Poly-Enrich results file for K562 NRSF. Most results where both methods are significant were for proximity to enhancers, similar to those of Gm12878 NRSF.

Supplementary Table 3.6: A table showing the number of peaks in the promoter region of genes in the true positive set, compared to the total number assigned to the genes. The proportion of promoter peaks in true positive genes is significantly greater than the proportion of promoter peaks not in the true positive genes.

CHAPTER 4 A Low-Rank Special Case in a Penalized Quasi-Likelihood Differential Expression Model

The content of chapter is planned to be submitted as an application note in Bioinformatics.

Introduction

Studying the transcriptome, or an organism's set of RNA molecules, allows understanding of genetic processes beyond the genome. Even if a gene is present in an organism, it may only expressed in certain biological contexts. For example, you would expect cell cycle genes to be expressed during stages of growth of development. Knowing which genes are expressed, especially during critical times such as carcinogenesis, provide a fuller understanding and enable treatments such as targeted gene treatments (Wirth and Ylä-Herttuala 2014).

Differential expression (DE) analysis allows investigators to identify genes that are over- or under-expressed in one context compared to another. Traditionally, investigators would perform RNA-seq on bulk tissue to obtain gene expression data in contexts of their choosing. However, this data would come from an aggregate of all cells in the tissue, which would consist of many different types of cells. Thus, single-cell RNA-seq (scRNA-seq) was invented to separate the expression profiles for individual cells in the experiment (Eberwine et al. 2014). This allows investigators to examine individual types of cells instead of the sum of all cells, providing opportunities to answer previously impossible questions, such as the differences between types of cells. However, because the gene expression profile is now divided among up to several thousand cells, the per-cell counts are much sparser, making it critical to model count data directly. Besides count data modeling,

two additional complications exist. First, certain scRNA-seq technologies (primarily non unique molecular identifier based ones) often produce many zero reads counts as technical 'dropouts' that should not actually be zeroes (Goeman and Bühlmann 2007). Second, some individual cells may become outliers with surprisingly large amounts of gene expression level, which is either caused by transcriptional bursting (Larsson et al. 2019) or accidental reads from contamination from other cells (McGinnis, Murrow, and Gartner 2019). Nevertheless, it has been recently shown that differential expression methods designed for bulk RNA-seq data can also be applied to analyze single-cell RNA-seq data, especially for single-cell RNA-seq data collected based on unique molecular identifier based technologies (Soneson and Robinson 2018).

Some differential expression analysis for single-cell RNA-seq expression data require cell types to be first identified, often through cell type clustering. Cell type clustering in the single cell data requires a first step of dimension reduction, relying on methods such as a t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton 2008) or uniform manifold approximation and projection (UMAP) (McInnes et al. 2018) to reduce the high-dimensional gene expression matrix into a low dimensional space. Afterwards, clustering algorithms such as the k-means algorithm are performed on the low dimensional space to clustering cells into cell clusters. Finally, by mapping known marker genes that can reliably identify cell types, each cluster can be assigned a cell type (Satija et al. 2015). Cell types can be types of differentiated cells (e.g. neuron, epithelial) or can be stages of development. The identified cell types may be used as categorical covariates themselves for comparing between cell types or used to stratify the data to test another predictor within a single cell type.

A common experimental design for differential expression analysis in single-cell RNA-seq studies is the hierarchical design, where there are several hundred to thousands of cells per individual, and individuals are categorized in comparing groups (e.g. cases and controls). Ideally, to identify genes that are differentially

expressed between groups, one would fit a mixed effects model with the group label to serve as the predictor variable and a random intercept term to represent individual-specific effects. Such mixed effects modeling requires the adaptation of generalized linear mixed models (GLMMs). However, algorithms for GLMMs tend to be computationally intensive, as it requires solving a high-dimensional integration that does not have an analytic solution. Indeed, obtaining results for a sample size in the tens of thousands like in the scale of scRNA-seq would require an infeasible amount of time. Previous workarounds include avoiding the use of GLMMs and instead of using a nested fixed effects model to adjust for each individual effect (Bakulski et al.). However, most such previous analyses ignore the hierarchical experimental design structure entirely and use a naïve two-sample approach, which can cause a bias if there is an individual with a comparatively large number of cells.

Here, we introduce a new computationally efficient algorithm that is particularly suitable for fitting GLMMs under the common hierarchical structure of scRNA-seq experimental design. We show that the resulting new method is well controlled for type I error, and is more powerful than other previous approaches of differential expression analysis under hierarchical design structure. Importantly, our method does not require astronomical computation time that are needed by fitting GLMMs using previous algorithms.

Methods

Model for Differential Expression

For a m by n gene by cell matrix with p individuals, we analyze each gene separately. For each gene, we take the row of gene counts for all cells $\boldsymbol{Y}$ and model this as:

$$Y_i \sim \text{Poisson}(\lambda_i)$$
$$\ln(\lambda_i) = \ln(T_i) + \mu + X_i\beta + u_j + \epsilon_i$$

where $i = 1,2,..,n$ is the cell index, T_i is the library size (sum of all counts over all genes) for the cell, X_i are the predictors and other covariates relating to the cell, u_j for $j = 1,2,\dots,p$ is the random individual effect with $\boldsymbol{u} \sim MVN_p(0, \sigma_u^2)$, and ϵ_i is the residual error with $\boldsymbol{\epsilon} \sim MVN_n(0, \sigma_\epsilon^2)$.

We use the algorithm in PQLseq as the basis of differential expression. PQLseq uses a penalized quasi-likelihood and iteratively re-weighted least squares to estimate the parameters. Derivation of the iteration steps can be seen in the Supplementary Material of the PQLseq paper (Sun et al. 2019).

Using low rank assumptions to decrease time complexity

While estimating the maximum quasi-likelihood, one of the required steps is to invert a *n* by *n* matrix, where *n* can be on the order of tens of thousands. Not only is matrix inversion extremely slow with complexity $O(n^3)$ without any assumptions, this is also required in every iteration, which further slows down the algorithm. However, when we define a *n* by *p* matrix $\boldsymbol{z}$ such that $z_{ik} = 1$ if cell *i* belongs to individual *k*, we can define the *n* by *n* kinship matrix used in PQLseq as $\boldsymbol{K} = \boldsymbol{z}^T\boldsymbol{z}$, which is of rank $p << n$. We can then use properties of low rank matrices to save computation time.

In the IRWLS algorithm, the matrix that needs to be inverted is (Equation 5 in PQLseq supplement)

$$\boldsymbol{H} = \boldsymbol{D^{-1}} + \sigma_u^2\boldsymbol{z}\boldsymbol{z^T} + \sigma_\epsilon^2\boldsymbol{I_n}$$

where $\boldsymbol{D}$ is a diagonal matrix. If we define $\boldsymbol{A}$ as

$$\boldsymbol{A} = \boldsymbol{D^{-1}} + \sigma_\epsilon^2\boldsymbol{I_p}$$

then $\boldsymbol{A}$ is also a diagonal matrix and we can use Woodbury's matrix identity (Woodbury 1950) to invert $\boldsymbol{H} = \boldsymbol{A} + \sigma_u^2\boldsymbol{z}\boldsymbol{z^T}$:

$$\boldsymbol{H^{-1}} = \boldsymbol{A^{-1}} - \boldsymbol{A^{-1}z}\left(\sigma_u^{-2}\boldsymbol{I_p} + \boldsymbol{z^TA^{-1}z}\right)^{-1}\boldsymbol{z^TA^{-1}}$$

There is still a required matrix inversion step, however it is for a p by p matrix instead with complexity $O(p^3)$, which is negligible compared to the next largest operation of matrix multiplication between n by n and n by p matrices, having complexity $O(\sim n^2)$.

Other differential expression methods

We compare our method with linear regression, bulk RNA-seq DE methods edgeR (Robinson, McCarthy, and Smyth 2010), DESeq2 (Love, Huber, and Anders 2014), and limma-voom (Law et al. 2014), and scRNA-seq DE method MAST (Finak et al. 2015). For each method, we use the same adjustments and options as the highest ranked evaluation given by Soneson et al (Soneson and Robinson 2018), which are explained in detail in the individual method descriptions below. Additionally, if a method is able to use multiple covariates in its design, we used the nested fixed effects design matrix that imitates a mixed effects model. Additional details for each method individually are listed below.

We use a plain multiple linear regression model as a "control" model, including only the nested design matrix and the logit of the cell's detection rate (proportion of nonzero counts) as recommended by Soneseon et al (Soneson and Robinson 2018). This design matrix will also be used in other methods that allow for multiple covariates. We use the default *lm* function in R to run the regression.

EdgeR (Robinson, McCarthy, and Smyth 2010) uses a negative binomial model originally designed for DE in bulk RNA-seq count data. As recommended by Van den Berge et al (Van den Berge et al. 2018), in order to better accommodate for zero inflation in scRNA-seq counts, we first use zingeR to generate observation weights to be used in edgeR. This method can handle multiple covariates so we include the nested design and the logit of each cell's detection rate. We test for the covariate of interest using the *glmQLFfit* function. Finally, we included a version where we aggregated the counts in each cell per individual to imitate a bulk RNA-seq experiment.

DESeq2 (Love, Huber, and Anders 2014) also uses a negative binomial distribution and multiple covariates. Its main difference from edgeR is that it uses a different strategy to normalize each cell. This method also allows for observation weights, which means zingeR is used to generate weights for for zero-inflation. The *nbinomWaldTest* function was used to test for the covariate of interest.

Limma-voom (Law et al. 2014) models the counts with a linear model and models the predicted variances to generate weights based on the expected mean-variance trend of a negative binomial model. We used the *voom* function to generate the weights and the *lmFit* function to estimate the coefficients.

MAST (Finak et al. 2015) uses a two-part model: a logistic regression on the dropout rate and a linear model on the counts given the cell is not a dropout. We use the *zlm* function for fitting the model and *lrTest* to test for the significance of the covariate of interest.

All calculations were performed using R version 3.6.1.

Real scRNA-seq data

The real scRNA-seq dataset used to evaluate the method is from an experiment performed on 8 African Americans and 7 European Americans on myoepithelial and luminal cells (Thong and Colacino). This data has 33,794 genes and 13,161 cells, with 6,847 of the cells being myoepithelial cells and the remaining 6,314 being luminal cells. Analysis was done by filtering the cells to only include one type of cell and testing for differential expression between African Americans vs European Americans.

Pre-analysis filtering

We first filtered out all genes with fewer than 5% positive counts among all cells, resulting in 5464 remaining genes. Lowering the number of genes analyzed lowers the total runtime of each method, with the amount saved dependent on the method. For most methods, including ours, reducing the proportion of genes analyzed linearly affects the computation time (i.e. 50% fewer genes results in approximately 50% less computation time).

Simulations imitating real data values

Starting with an analysis of the myoepithelial cells, we obtain a distribution of parameter estimates in the model. We used the median of the estimates from all genes for $\mu, \sigma_u, \sigma_\epsilon$ as the basis for our model-based simulations. We then freely control β to simulate several signal strengths.

Simulations for Type I error and Statistical Power

We simulate the null hypothesis of "no differential expression between groups" in the following: Given cell sequencing depths T and a single cell group indicator X; fix parameters $\mu, \sigma_u, \sigma_\epsilon$ and set $\beta = 0$ to generate a random counts matrix using the model for a large number of genes. We also generate a second type of null data using a real scRNA-seq experiment with a known group effect for differential expression and randomize the group assignment variable.

Evaluating Type I error

Type I error for a method is evaluated using a Q-Q plot, plotting the $-\log_{10}$ p-values of the methods run on the simulated null distributions mentioned above, against a uniformly distributed set of p-values. All methods use unadjusted p-values directly from the output and no other modifications.

Evaluating Statistical Power

To ensure that methods with inflated p-values do not automatically have larger power from detecting true positives, we use an FDR-controlled significance threshold. We first simulate an empirical null distribution of p-values by analyzing a

null simulation ten times. Then, for a given threshold, we find the p-value in the permutations such that its quantile matches the threshold and use that value as the FDR-controlled cutoff.

Results

Using the low-rank property significantly decreases computation time

Using Woodbury's matrix identity (Woodbury 1950) to reduce the problem of inverting a $n \times n$ matrix into a $p \times p$ matrix significantly decreases the analysis time. We simulated several genes by varying different sizes of *n* and *p* to illustrate the effect of size on computation time. The times can be seen in Table 4.1. We see that increasing the number of total cells increase computation time as expected, and increasing number of individuals does not affect it much, however the low-rank version is able to finish over ten times faster when there are only 1000 total cells, which is already much smaller than most real scRNA-seq sample sizes. We can imagine that trying the original version may take months to run a data set in the expected size of 10,000s of cells and thousands of genes.

n/p	Low rank	Original version
100/4	0.329	0.510
200/8	0.449	1.12
400/4	0.838	4.873
400/8	0.857	5.006
1000/4	5.782	61.34
1000/8	5.545	59.306
1000/20	6.430	59.406
2000/10	26.332	1232.739

Table 4.1 Time, in seconds, to analyze 10 simulated genes with a data set of *n* cells divided evenly amongst *p* individuals.

We see the time increases when *n* increases while changes in *p* are negligible, but the low rank version is much faster than the original version as *n* increases.

Type I error simulations show controlled errors

Starting with the base estimates of $\mu = -7, \sigma_u = 0.25, \sigma_\epsilon = 1$, we then used $\beta = 0$ to simulate a null distribution. We ran all methods on this simulated data to generate a distribution of p-values. The Q-Q plot of p-values can be shown in Figure 4.1. We see that our method is the most conservative out of all methods. Linear regression, aggregated EdgeR, and MAST are also conservative while EdgeR, limma, and DESeq2 are extremely inflated.

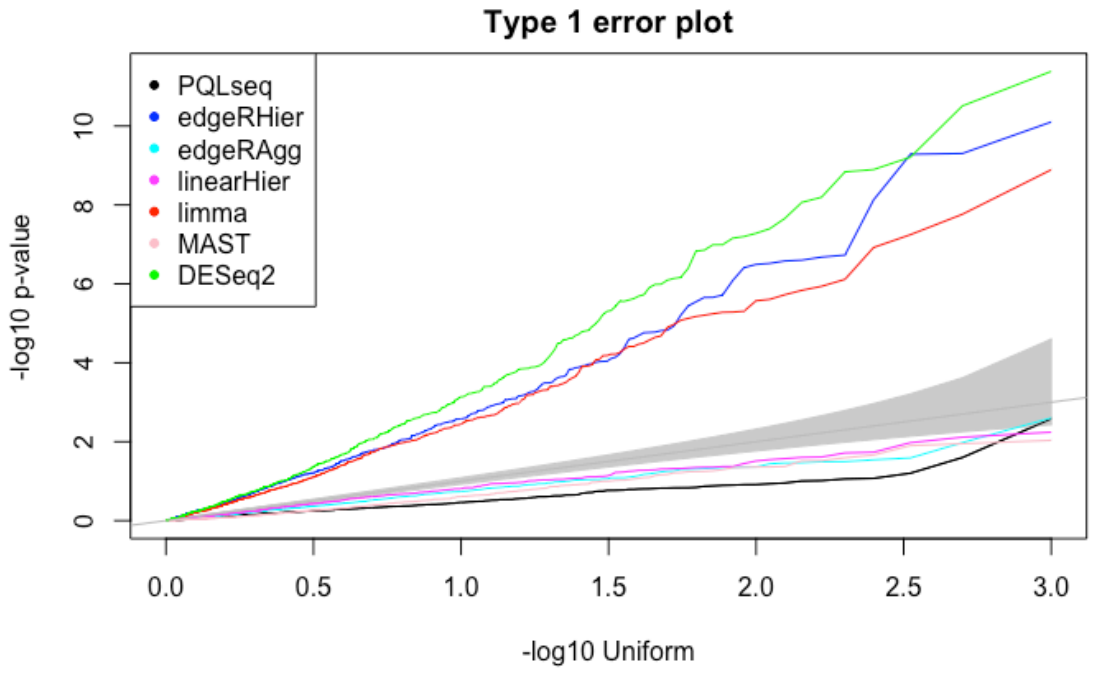


Figure 4.1 Q-Q plots of p-values for each method on the null simulation of setting $\beta = 0$.

DESeq2, edgeR, and limma have greatly inflated p-values while aggregated edgeR, linear regression, MAST, and PQLseq have slightly deflated p-values.

Power simulations show PQLseq is most powerful

We used the simulated data of $\beta = 0$ as an empirical null distribution to make an FDR-controlled threshold for each method. We then chose β as 0.25,0.5,0.75, and 1 to simulate. The power graph for FDR-thresholds of 0.05 and 0.001 can be seen in Figure 4.2. We see that our method is the most powerful. We also see several

methods with a uniform zero power, which are the same methods as the ones with inflated Type I error in Figure 4.1, except for aggregated edgeR.

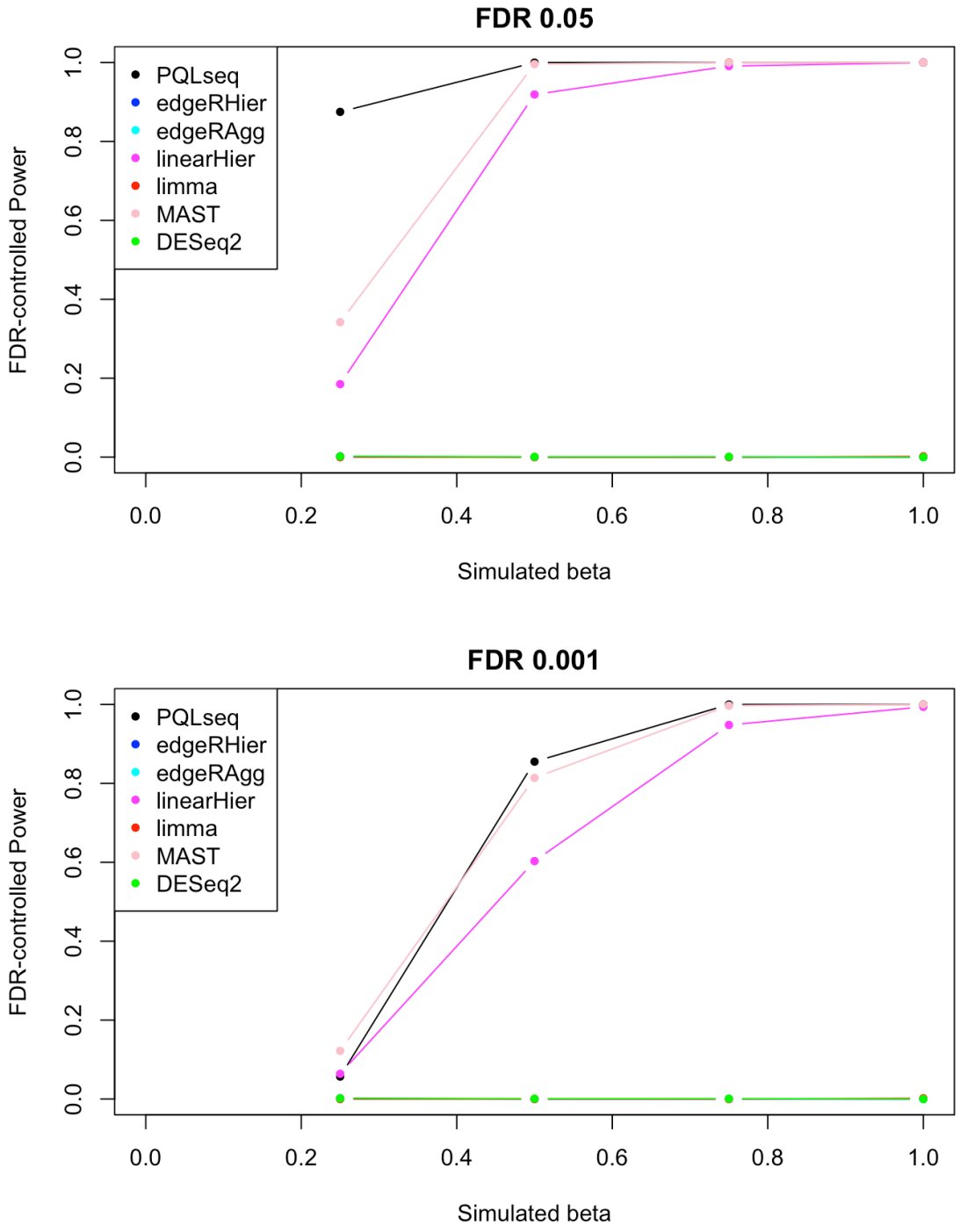


Figure 4.2 Power simulations for several levels of betas: 0.25,0.5,0.75, and 1 for FDR cutoffs of 0.05 and 0.001.

We see that our method has the highest power out of all methods for the FDR cutoff of 0.05. MAST outperforms slightly at an FDR cutoff of 0.001 for β-0.25, which is likely due to PQLseq having less conservative low-end p-values (see Figure 4.1). We see several methods have near 0 power for all simulated betas, which are also the methods that have an inflated Type I error from Figure 4.1 except for aggregated edgeR.

Simulations varying other parameters show consistent behavior

We also varied the other parameters to test how they affect the power. We start with initial values of $\beta = 0.25, \mu = -7, \sigma_u^2 = 0.25, \sigma_\epsilon^2 = 1$, which has a power of 0.875, and modify μ, σ_u^2, and σ_ϵ^2 one at a time. The powers can be seen in Table 4.2. We see that a lower μ decreases power, which would simulate significantly more zeroes, making it more difficult to detect the signal. We also see that increasing either σ decreases power, which is consistent with the concept of higher variance being more difficult to model. We also see that increasing the variance of $\boldsymbol{u}$, the individual random effect, significantly reduces power rapidly. It is likely that such a high variance combined with a low individual count, when compared to the effect size of only 0.25, greatly hinders the simulated data to generate a correct effect size.

$\boldsymbol{\mu}$	**-10**	***-7***	**-4**	
Power	0.446	*0.875*	0.917	
$\boldsymbol{\sigma_u^2}$	**0.1**	***0.25***	**0.33**	**0.5**
Power	0.967	*0.875*	0.608	0.006
$\boldsymbol{\sigma_e^2}$	**0.25**	**0.5**	***1.0***	**1.5**
Power	0.997	0.991	*0.875*	0.677

Table 4.2 Power for PQLseq when modifying other variables.

Starting with the initial values of $\beta = 0.25, \mu = -7, \sigma_u^2 = 0.25, \sigma_\epsilon^2 = 1$. The values in italics are the values that were initially chosen in the previous simulations. Decreasing μ or increasing σ_u^2 or σ_ϵ^2 reduces power as expected.

PQLseq is robust to outliers

To simulate an outlier, we randomly picked a single cell and multiplied all of its counts by 2 for a realistic scenario and 10 for an exaggerated scenario. We then ran the method on this data as well as the same data with the outlier removed. We see

that multiplying an entire cell's counts make very little effect on the overall analysis (Supplementary Figure 4.1). We confirm that the library size can properly account for high outlier counts.

Real data permutations show PQLseq is well controlled

With the randomized data subsetted to only the myoepithelial cell type, we simulated the null by permuting the individual assignment labels. For each method, we did three permutations. The Q-Q plots for the p-values can be seen in Figure 4.3.

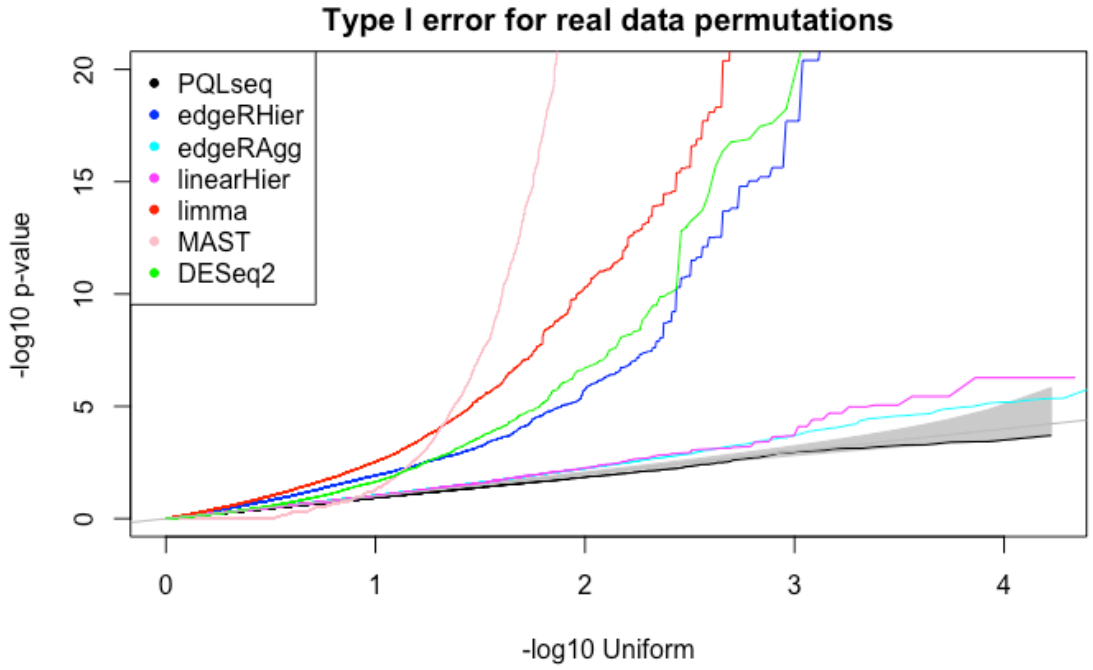


Figure 4.3 Q-Q plots for all methods on randomized real data.

We see PQLseq is the most well-controlled with aggregated edgeR and linear regression being the next best. The other four methods are massively inflated.

PQLseq detects the most DE genes in real data

Using the same strategy as the simulated data, we measure the power of each method, which is listed in Table 4.3. Although all methods detected few DE genes, which is likely due to the real data not having many true DE genes to begin with,

PQLseq still found the most DE genes given an FDR cutoff of 0.05. The methods with zero significant results (limma, MAST, DESeq2) also have inflated Type I errors (Figure 4.3) and thus would require a much lower p-value for actual significance. For limma, MAST, and DESeq2, using a Bonferroni-Hochberg FDR adjustment results in 5.7%, 24.2%, and 5.6% significant DE genes, respectively.

Method	PQLseq	edgeR	edgeR (agg)	Lin.reg	limma	MAST	DESeq2
Power ($x10^{-3}$)	2.76	0.02	0.01	0.11	0	0	0

Table 4.3 Proportion of significantly DE genes in myoepithelial cells.

PQLseq finds a much higher proportion of DE genes compared to other methods. The overall low amount of significant findings is likely due to the data itself not having many true DE genes.

Availability and Usage

The method can be accessed through the Github respository: https://github.com/leetaiyi/PQLseq, which will be pushed to the master Github: https://github.com/sqsun/PQLseq and CRAN when finalized. The *pqlseq* function has an added parameter, *lowrank*, which should be set to *TRUE* and the *RelatednessMatrix* parameter set to the *n* by *p* indicator matrix $\boldsymbol{z}$ where $z_{ik} = 1$ if cell *i* belongs to individual *k*. This is equivalent to setting *lowrank* to be *FALSE* and setting *RelatednessMatrix* to be $\boldsymbol{z}^T\boldsymbol{z}$, but is several magnitudes faster.

Code used to generate results for this chapter can be found in https://github.com/leetaiyi/Dissertation/tree/master/Chapter%204

Discussion

The common hierarchical data structures have historically not been adjusted for while performing scRNA-seq data analysis, with one of the largest reasons being that methods to do so have time complexities. By taking advantage of the hierarchical data structure, we greatly decreased the time complexity of PQLseq's

estimating algorithm. This creates an opportunity for investigators to properly adjust for the hierarchical data structure, which becomes extremely important in scenarios where one individual has many more measured cells than any other.

We showed that using a random effects model to adjust for individual effects in a hierarchical data structure increases power for the analysis without impacting the Type 1 error in both simulations and in a real data example. It is unknown if the detected DE genes in the myoepithelial cells are actually truly DE in the context of African vs. European Americans as there are not many sources of true positives.

Despite the large time save, PQLseq still takes a longer time to run compared to methods without random effects. The next possible avenues to reduce computation time are speeding up matrix multiplications or utilizing a faster converging algorithm for maximum likelihood estimation.

Our new method is shown to perform well in hierarchical data structures, but it would likely not perform as well as the other methods if the data more closely follows different modeling assumptions. Thus it is important for an investigator to know what methods best fit their data, and there will likely never be a method that is uniformly superior to all other methods.

Future Ideas

Several models attempt to account for zero-inflation by also modeling the probability of a dropout given the covariates of interest. An additional assumption is that the higher the mean of the cell counts, the less likely it is for that cell to have dropouts and any zeroes are more likely to be true zeroes. Currently, all models that also model the dropout have a parameter independent to the parameter that models the mean counts. The idea is to couple the estimated parameters in a way such that an increase in the counts estimate also increases the dropout rate estimate. A proposed model is below:

For cell $i = 1..n$,

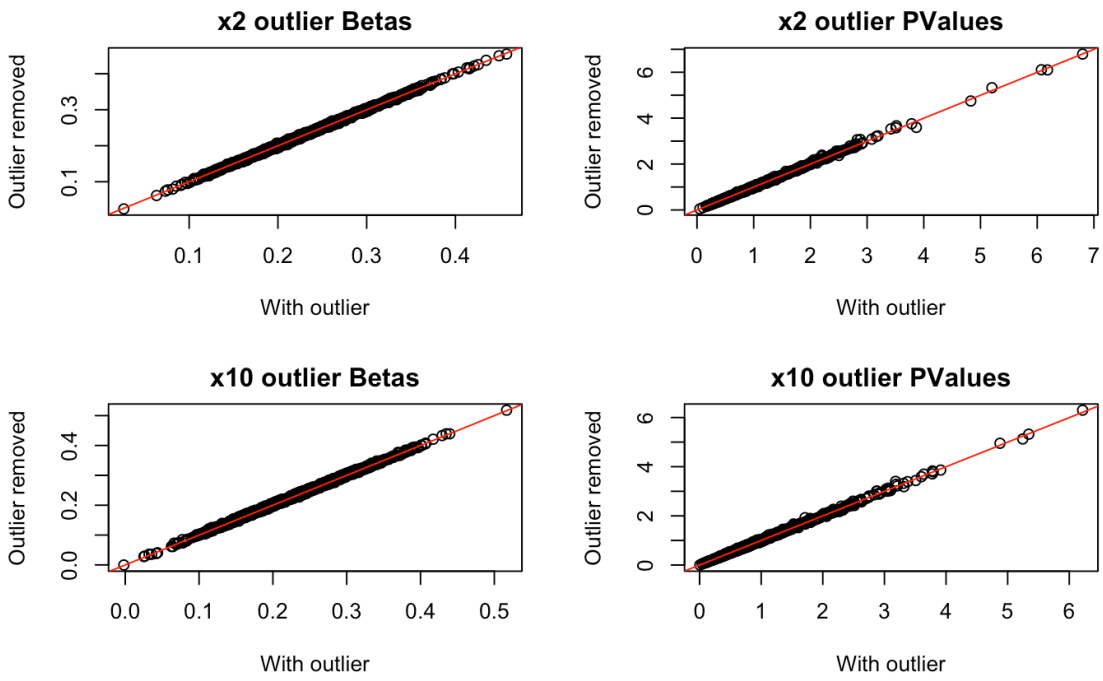
$$Y_i \sim \pi_i Poi(\lambda_i) + (1 - \pi_i)\delta_0$$

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \mu_1 + \boldsymbol{X}_{i\cdot}(\alpha \circ \beta) + \gamma Z_i$$

$$\ln(\lambda) = \ln(T_i) + \mu_2 + \boldsymbol{X}_{i\cdot}\beta + e_i$$

where δ_0 is a point density at 0, $\boldsymbol{X}_{i\cdot}$ is a vector of covariates, Z_i is a covariate specific to dropout rate (e.g. detection rate), T_i is the library size, and $e_i \sim \ln Gamma(\psi, \psi)$. An attempted implementation of this model can be found at: https://github.com/leetaiyi/scPoissonGamma

Supplementary Figures for Chapter 4



Supplementary Figure 4.1
Data simulations with one cell's counts multiplied by 2 and by 10 to simulate an outlier. We ran our method with and without the outlier and compared their results. We see an almost perfect correlation, indicating that the outlier does not affect the results, so the method is robust to outliers.

Chapter 5 Discussion

Summary

With the new methods introduced in the above chapters, we have opened up several options to analyze new types of gene regulation and expression data. A count-based approach for genomic regions allows another option when a binary approach fails. A different perspective in proximal binding to regulatory regions shows more insight in binding behavior. And finally, the ability to use a common data structure in scRNA-seq without requiring impractical runtimes allows for more investigators to run more powerful analyses. Below, I review the motivations, development, hardships, and future paths of the dissertation projects.

Poly-Enrich enables opportunities in other types of genomic regions

Poly-Enrich allows for gene set enrichment analysis of many types of genomic regions with the assumption that gene regulation is incremental with more binding sites. It also allows for weights for each genomic region, which can be used to include relevant information such as peak binding strength, distance to a transcription start site, level of change in open chromatin, or adjusting for multiple gene assignments. For several of these, it remains to be assessed how well they improve GSE results.

When analyzing very large sets of genomic regions, such that almost all genes have at least one assigned genomic region, binary score methods such as ChIP-Enrich have trouble detecting any enrichment, whereas count-based methods such as Poly-Enrich retain their power. However, this does not mean Poly-Enrich is uniformly superior, as a binary score may actually better model some regulatory mechanics, an example scenario being one protein binding site blocking transcription, while

additional binding events have no additional effect. Currently, it is not obvious how most TFs regulate a gene; that is, whether more binding sites implies incremental regulation, one binding site is functionally the same as many, a broader binding site implies more regulation, or some other association. Thus, it is important to have several modeling options available in GSE tests for various types of genomic regions.

Even with several GSE testing options, there may not be an obvious "correct" method for a specific analysis. The hybrid method allows one to combine several GSE methods in case the situation is not obvious. While the current hybrid method is controlled for Type I error, it is a conservative approach, and it is possible to derive a more powerful hybrid method by using the properties of the tests, for example the binomial and negative binomial distributions for ChIP- and Poly-Enrich. However, this will not always be possible given arbitrary GSE tests, and it will be up to the investigator to decide on the best strategy for their particular data.

ProxReg gives another perspective to pathway regulation

Most GSE methods may have positive results that, despite being true positives in the context of the method, have other missing details due to its design. For instance, Poly-Enrich, which tests for number of peaks, can detect if genes in a gene set have more binding events, but not if those binding events are truly regulating the genes due to enhancer or promoter binding. Thus, it is important to look at other available perspectives, as proxReg's test for proximal binding to regulatory regions serves as a complement to GSE methods. ProxReg was able to confirm true positives and find false positives from Poly-Enrich results based on the proximal binding to TSSes.

A future idea is to consider the integration between a ChIP-seq experiment and a related RNA-seq experiment (e.g. comparing cells that are wild type versus knockout for the TF used in the ChIP-seq experiment). Given a biological context, ChIP-seq data gives information on where a TF binds but not necessarily its effects, while the RNA-seq data reveals the changes in gene expression, but not information

regarding whether their regulation is direct versus indirect (i.e. secondary or tertiary downstream effects). However, if we combine both ChIP-seq and RNA-seq data from the same context, we may conclude that: a gene set is truly regulated by a TF if both data have signal; a gene set is regulated by the TF but not in the context if ChIP-seq has a signal but RNA-seq does not; a gene set is relevant in the context but not regulated by the TF if RNA-seq has signal but ChIP-seq does not; and possibly irrelevant if neither are significant. This could be modeled, for example, by using a logistic model for gene set inclusion with ChIP-seq peak binding and RNA-seq differential expression as covariates. This idea combines two perspectives, similar to how proxReg adds another perspective, to provide a larger picture of a TF's role in gene regulation.

Creating more accurate distal regulatory element locus definitions

ProxReg also found TFs that bind to genes in certain gene sets far from promoter regions but near distal regulatory regions instead. However, the peak-to-gene assignments in proxReg are still currently based on their nearest gene TSS. If we instead use the new enhancer locus definition for gene assignments, we hypothesized that it would improve the test by reducing the number of misassigned peaks, allowing us to more properly assess a TF's role in a pathway. Poly-Enrich was helpful in evaluating a large set of DRE locus definitions to find the one that was able to most accurately identify true positive gene sets. This top ranked enhancer locus definition, along with 18 others, performed significantly better than simply assigning peaks to the gene with the nearest TSS. These enhancer locations can then be used in any GSE test that requires assigning peaks to genes, and it will greatly improve the test as more peaks will be correctly assigned to their target gene compared to naïvely assigned every peak to their nearest gene TSS.

Faster implementation of scRNA-seq data analysis for hierarchical data structures

In the context of scRNA differential expression, we modified PQLseq, an existing method capable of performing random effects models to be able to run it in the scope of larger sample sizes often seen in scRNA-seq, without requiring astronomical computation times. With the hierarchical data structure being extremely common in scRNA-seq experiments, methods that can properly utilize its assumptions will be very helpful for more powerful analyses when compared to naïve methods that ignore assumptions. We showed that PQLseq increases power when compared to other popular differential expression methods in hierarchical data structures. This method will be a useful tool for better differential expression analysis and has already been utilized in a study of breast cancer, where it showed differential expression of cancer and embryonic stem cell genes in cancerous mammary mouse cells (Thong et al.).

Future ideas in scRNA-seq

One of the intermediate steps in scRNA-seq analysis is to cluster cells with similar expression profiles to be able to identify cell types. However, cluster assignments tend to be an all-or-nothing assignment, which makes less sense for cells that are clustered near the boundaries of clusters. One idea of a workaround is to assign cluster weights to cells based on a confidence level, similar to MAGIC (van Dijk et al. 2018), and then use these weights instead of separating analyses by cell type. This could be achieved by instead of using indicators for each cell type, use the weights (e.g. the probability of being the cell type) as the covariates instead.

As yet, there is no standard approach for performing pathway analysis on single cell transcriptomics data. One approach that's been proposed is to skip the cluster assignments and instead perform pathway analysis directly after gene quantification. One approach, PAGODA (Fan et al. 2016), achieves this by using PCA on the genes in a gene set and testing if the magnitude of the first eigenvalue is significant. One novel idea is to perform pathway analysis based on the tSNE or

UMAP (McInnes et al. 2018) plot of the data. One could compare the expression "temperature map" between genes to create a gene-by-gene correlation matrix using some metric of similarity between the visualization plots (Figure 5.1). Then, we would use a gene set assignment to determine the variance effect of this correlation matrix to find significant gene sets. Such an idea is explored by Cho et al (Cho et al. 2019), where they obtain a gene-gene correlation matrix using a bivariate zero-inflated negative binomial model, but have yet to find a method to analyze the resulting matrix. Additionally, there are new developments of assays that can not only measure the gene expression level of each cell, but do so while keeping the tissue intact to obtain a spatial landscape of cells. Currently there are very few methods to account for the spatial location of cells (Svensson, Teichmann, and Stegle 2018);(Achim et al. 2015);(Baccin et al. 2019). Similar to the approach of comparing temperature maps of visualization plots, one could instead compare spatial counts between genes to create a correlation matrix.

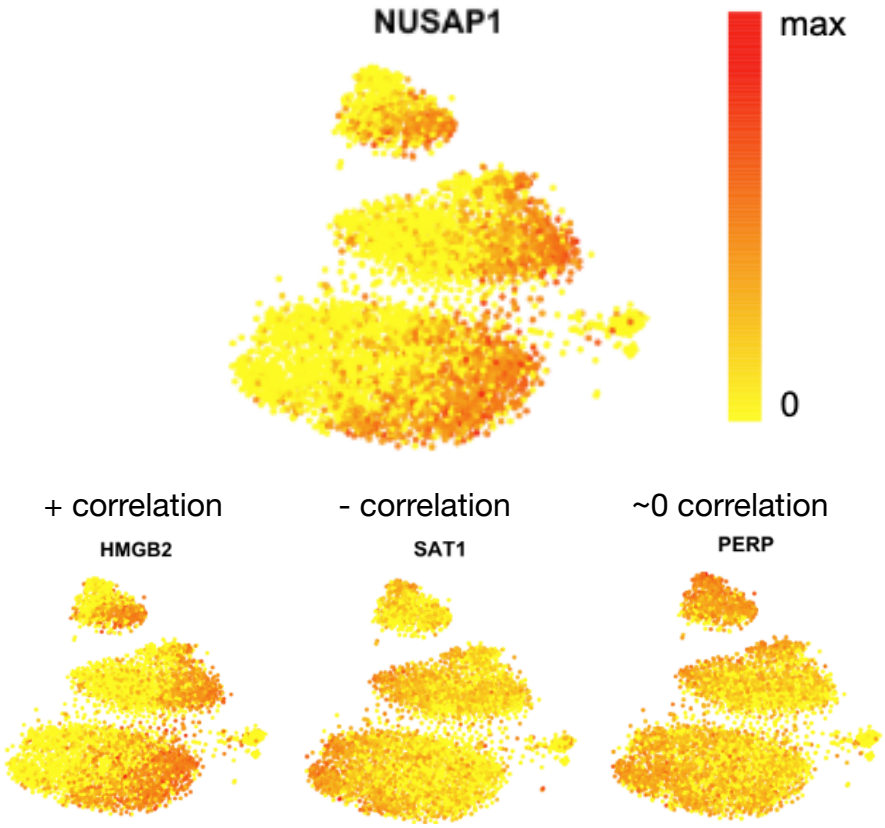


Figure 5.1 The framework for using scRNA-seq temperature maps to pairwise compare genes.

The goal is to have a gene-gene correlation matrix based on some metric that can compares the patterns between two maps.

Closing statements

Biological organisms are the product of random, brute-force approaches to survival in combination with natural selection spanning billions of years. Similar to how an inventor of a machine-learning algorithm cannot explain the resulting creation, it is also equivalently difficult to explain how the genetic code and its interpreter came to exist in its current form. The genome is just one part of the story; its regulation is an equally important part. With the invention of new technologies that reveal more to the intricacies of genetics such as ChIP-seq and single-cell RNA-seq, methods are required to be able to interpret and analyze their data. There will inevitably be more discoveries to allow for different perspectives to understanding biology, and the opportunities for more statistical methods will increase yet again.

References

Achim, K., J. B. Pettit, L. R. Saraiva, D. Gavriouchkina, T. Larsson, D. Arendt, and J. C. Marioni. 2015. "High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin." *Nat Biotechnol* 33 (5):503-9. doi: 10.1038/nbt.3209.

Aldrup-Macdonald, M. E., and B. A. Sullivan. 2014. "The past, present, and future of human centromere genomics." *Genes (Basel)* 5 (1):33-50.

Alhamdoosh, M., C. W. Law, L. Tian, J. M. Sheridan, M. Ng, and M. E. Ritchie. 2017. "Easy and efficient ensemble gene set testing with EGSEA." *F1000Res* 6:2010. doi: 10.12688/f1000research.12544.1.

Allocco, D. J., I. S. Kohane, and A. J. Butte. 2004. "Quantifying the relationship between co-expression, co-regulation and gene function." *BMC Bioinformatics* 5:18. doi: 10.1186/1471-2105-5-18.

Andersson, Robin. 2015. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. Bioessays.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nat Genet* 25 (1):25-9. doi: 10.1038/75556.

Baccin, Chiara, Jude Al-Sabah, Lars Velten, Patrick M. Helbling, Florian Grünschläger, Pablo Hernández-Malmierca, César Nombela-Arrieta, Lars M. Steinmetz, Andreas Trumpp, and Simon Haas. 2019. "Combined single-cell and spatial transcriptomics reveals the molecular, cellular and spatial bone marrow niche organization." *bioRxiv*:718395. doi: 10.1101/718395.

Bais, A. S., and D. Kostka. 2019. "scds: Computational Annotation of Doublets in Single-Cell RNA Sequencing Data." *Bioinformatics*. doi: 10.1093/bioinformatics/btz698.

Bakulski, Kelly M, John F Dou, Robert C Thompson, Christopher T Lee, Lauren Y Middleton, Bambarendage P U Perera, Sean P Ferris, Tamara R Jones, Kari Neier, Xiang Zhou, Maureen A Sartor, Saher S Hammoud, Dana C Dolinoy, and Justin A Colacino. Mapping the effects of developmental lead (Pb) exposure on the hippocampus with single cell analysis. In preparation.

Benjamini, Yoav and Hochberg, Yosef. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing}. Journal of the Royal statistical society: series B (Methodological).

Berger, S. L. 2007. "The complex language of chromatin regulation during transcription." *Nature* 447 (7143):407-12. doi: 10.1038/nature05915.
Boyle, A. P., S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford. 2008. "High-resolution mapping and characterization of open chromatin across the genome." *Cell* 132 (2):311-22. doi: 10.1016/j.cell.2007.12.014.
Brunner, A. M., J. C. Schimenti, and C. H. Duncan. 1986. "Dual evolutionary modes in the bovine globin locus." *Biochemistry* 25 (18):5028-35.
Buenrostro, J. D., B. Wu, H. Y. Chang, and W. J. Greenleaf. 2015. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide." *Curr Protoc Mol Biol* 109:21.29.1-21.29.9. doi: 10.1002/0471142727.mb2129s109.
Cai, Wenqing, Huang, Jialiang, Zhu, Qian, Li, E. Bin, Seruggia, David, Zhou, Pingzhu, Nguyen, Minh, Fujiwara, Yuko, Xie, Huafeng, and Yang. 2019. Enhancer-dependance of gene expression increases with developmental age. bioRxiv.
Carlson, M, and B Maintainer. 2015. "TxDb. Hsapiens. UCSC. hg19. knownGene: Annotation package for TxDb object (s)." In *R package version 3.0.0*.
Carlson, Marc. 2018. org.Hs.eg.db: Genome wide annotation for Human.
Carlson, Marc. 2019. GO.db: A set of annotation maps describing the entire Gene Ontology. R Package.
Cavalcante, R. G., C. Lee, R. P. Welch, S. Patil, T. Weymouth, L. J. Scott, and M. A. Sartor. 2014. "Broad-Enrich: functional interpretation of large sets of broad genomic regions." *Bioinformatics* 30 (17):i393-400. doi: 10.1093/bioinformatics/btu444.
Chen, J., Z. Hu, M. Phatak, J. Reichard, J. M. Freudenberg, S. Sivaganesan, and M. Medvedovic. 2013. "Genome-wide signatures of transcription factor activity: connecting transcription factors, disease, and small molecules." *PLoS Comput Biol* 9 (9):e1003198. doi: 10.1371/journal.pcbi.1003198.
Chicco, Davide, Haixin Sarah Bi, Jüri Reimand, and Michael M Hoffman. 2019. "BEHST: genomic set enrichment analysis enhanced through integration of chromatin long-range interactions." *bioRxiv* (168427).
Cho, Hunyong, Chuwen Liu, Jinyoung Park, and Di Wu. 2019. bzinb: Bivariate Zero-Inflated Negative Binomial Model Estimator. CRAN.
Chu, Y., and D. R. Corey. 2012. "RNA sequencing: platform selection, experimental design, and data interpretation." *Nucleic Acid Ther* 22 (4):271-4. doi: 10.1089/nat.2012.0367.
Consortium, ENCODE Project. 2004. "The ENCODE (ENCyclopedia Of DNA Elements) Project." *Science* 306 (5696):636-40. doi: 10.1126/science.1105136.
Cullen, E. M., J. C. Brazil, and C. M. O'Connor. 2010. "Mature human neutrophils constitutively express the transcription factor EGR-1." *Mol Immunol* 47 (9):1701-9. doi: 10.1016/j.molimm.2010.03.003.
D'haeseleer, P. 2006. "What are DNA sequence motifs?" *Nat Biotechnol* 24 (4):423-5. doi: 10.1038/nbt0406-423.
Davis, A. P., C. J. Grondin, R. J. Johnson, D. Sciaky, B. L. King, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly. 2017. "The Comparative Toxicogenomics Database: update 2017." *Nucleic Acids Res* 45 (D1):D972-D978. doi: 10.1093/nar/gkw838.

de Koning, A. P., W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock. 2011. "Repetitive elements may comprise over two-thirds of the human genome." *PLoS Genet* 7 (12):e1002384. doi: 10.1371/journal.pgen.1002384.

De Luca, A., A. Severino, P. De Paolis, G. Cottone, L. De Luca, M. De Falco, A. Porcellini, M. Volpe, and G. Condorelli. 2003. "p300/cAMP-response-element-binding-protein ('CREB')-binding protein (CBP) modulates co-operation between myocyte enhancer factor 2A (MEF2A) and thyroid hormone receptor-retinoid X receptor." *Biochem J* 369 (Pt 3):477-84. doi: 10.1042/BJ20020057.

de Sotero-Caio, C. G., D. C. Cabral-de-Mello, M. D. S. Calixto, G. T. Valente, C. Martins, V. Loreto, M. J. de Souza, and N. Santos. 2017. "Centromeric enrichment of LINE-1 retrotransposons and its significance for the chromosome evolution of Phyllostomid bats." *Chromosome Res* 25 (3-4):313-325. doi: 10.1007/s10577-017-9565-9.

Deaton, A. M., and A. Bird. 2011. "CpG islands and the regulation of transcription." *Genes Dev* 25 (10):1010-22. doi: 10.1101/gad.2037511.

deHaseth, P. L., M. L. Zupancic, and M. T. Record. 1998. "RNA polymerase-promoter interactions: the comings and goings of RNA polymerase." *J Bacteriol* 180 (12):3019-25.

Dong, X., and Z. Weng. 2013. "The correlation between histone modifications and gene expression." *Epigenomics* 5 (2):113-6. doi: 10.2217/epi.13.13.

Dowen, J. M., Z. P. Fan, D. Hnisz, G. Ren, B. J. Abraham, L. N. Zhang, A. S. Weintraub, J. Schujiers, T. I. Lee, K. Zhao, and R. A. Young. 2014. "Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes." *Cell* 159 (2):374-387. doi: 10.1016/j.cell.2014.09.030.

Eberwine, J., J. Y. Sul, T. Bartfai, and J. Kim. 2014. "The promise of single-cell sequencing." *Nat Methods* 11 (1):25-7. doi: 10.1038/nmeth.2769.

Ernst, J., and M. Kellis. 2012. "ChromHMM: automating chromatin-state discovery and characterization." *Nat Methods* 9 (3):215-6. doi: 10.1038/nmeth.1906.

Ertosun, M. G., F. Z. Hapil, and O. Osman Nidai. 2016. "E2F1 transcription factor and its impact on growth factor and cytokine signaling." *Cytokine Growth Factor Rev* 31:17-25. doi: 10.1016/j.cytogfr.2016.02.001.

Fan, J., N. Salathia, R. Liu, G. E. Kaeser, Y. C. Yung, J. L. Herman, F. Kaper, J. B. Fan, K. Zhang, J. Chun, and P. V. Kharchenko. 2016. "Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis." *Nat Methods* 13 (3):241-4. doi: 10.1038/nmeth.3734.

Finak, G., A. McDavid, M. Yajima, J. Deng, V. Gersuk, A. K. Shalek, C. K. Slichter, H. W. Miller, M. J. McElrath, M. Prlic, P. S. Linsley, and R. Gottardo. 2015. "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data." *Genome Biol* 16:278. doi: 10.1186/s13059-015-0844-5.

Fryer, C. J., E. Lamar, I. Turbachova, C. Kintner, and K. A. Jones. 2002. "Mastermind mediates chromatin-specific transcription and turnover of the Notch enhancer complex." *Genes Dev* 16 (11):1397-411. doi: 10.1101/gad.991602.

Gertz, J., T. E. Reddy, K. E. Varley, M. J. Garabedian, and R. M. Myers. 2012. "Genistein and bisphenol A exposure cause estrogen receptor 1 to bind thousands of

sites in a cell type-specific manner." *Genome Res* 22 (11):2153-62. doi: 10.1101/gr.135681.111.

Giner, Goknur, and Gordon Smyth. 2016. "statmod: Probability Calculations for the Inverse Gaussian Distribution." *The R Journal* 8. doi: 10.32614/RJ-2016-024.

Giorgetti, L., T. Siggers, G. Tiana, G. Caprara, S. Notarbartolo, T. Corona, M. Pasparakis, P. Milani, M. L. Bulyk, and G. Natoli. 2010. "Noncooperative interactions between transcription factors and clustered DNA binding sites enable graded transcriptional responses to environmental inputs." *Mol Cell* 37 (3):418-28. doi: 10.1016/j.molcel.2010.01.016.

Goeman, J. J., and P. Bühlmann. 2007. "Analyzing gene expression data in terms of gene sets: methodological issues." *Bioinformatics* 23 (8):980-7. doi: 10.1093/bioinformatics/btm051.

Gotea, V., A. Visel, J. M. Westlund, M. A. Nobrega, L. A. Pennacchio, and I. Ovcharenko. 2010. "Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers." *Genome Res* 20 (5):565-77. doi: 10.1101/gr.104471.109.

Haeussler, M., A. S. Zweig, C. Tyner, M. L. Speir, K. R. Rosenbloom, B. J. Raney, C. M. Lee, B. T. Lee, A. S. Hinrichs, J. N. Gonzalez, D. Gibson, M. Diekhans, H. Clawson, J. Casper, G. P. Barber, D. Haussler, R. M. Kuhn, and W. J. Kent. 2019. "The UCSC Genome Browser database: 2019 update." *Nucleic Acids Res* 47 (D1):D853-D858. doi: 10.1093/nar/gky1095.

Harris, M. A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White, and Gene Ontology Consortium. 2004. "The Gene Ontology (GO) database and informatics resource." *Nucleic Acids Res* 32 (Database issue):D258-61. doi: 10.1093/nar/gkh036.

Heinz, S., C. E. Romanoski, C. Benner, and C. K. Glass. 2015. "The selection and function of cell type-specific enhancers." *Nat Rev Mol Cell Biol* 16 (3):144-54. doi: 10.1038/nrm3949.

Hsu, F., W. J. Kent, H. Clawson, R. M. Kuhn, M. Diekhans, and D. Haussler. 2006. "The UCSC Known Genes." *Bioinformatics* 22 (9):1036-46. doi: 10.1093/bioinformatics/btl048.

Huang, D. W., B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. 2007. "The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists." *Genome Biol* 8 (9):R183. doi: 10.1186/gb-2007-8-9-r183.

Jabbari, K., and G. Bernardi. 2004. "Cytosine methylation and CpG, TpG (CpA) and TpA frequencies." *Gene* 333:143-9. doi: 10.1016/j.gene.2004.02.043.

Kanehisa, M., and S. Goto. 2000. "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic Acids Res* 28 (1):27-30. doi: 10.1093/nar/28.1.27.

Karin, M. 1990. "Too many transcription factors: positive and negative interactions." *New Biol* 2 (2):126-31.

Kharchenko, P. V., L. Silberstein, and D. T. Scadden. 2014. "Bayesian approach to single-cell differential expression analysis." *Nat Methods* 11 (7):740-2. doi: 10.1038/nmeth.2967.

Koneva, L. A., Y. Zhang, S. Virani, P. B. Hall, J. B. McHugh, D. B. Chepeha, G. T. Wolf, T. E. Carey, L. S. Rozek, and M. A. Sartor. 2018. "HPV Integration in HNSCC Correlates with Survival Outcomes, Immune Response Signatures, and Candidate Drivers." *Mol Cancer Res* 16 (1):90-102. doi: 10.1158/1541-7786.MCR-17-0153.

Korthauer, K. D., L. F. Chu, M. A. Newton, Y. Li, J. Thomson, R. Stewart, and C. Kendziorski. 2016. "A statistical approach for identifying differential distributions in single-cell RNA-seq experiments." *Genome Biol* 17 (1):222. doi: 10.1186/s13059-016-1077-y.

Kouzarides, A., and T. Bannister. 2011. Regulation of chromatin by histone modifications. Cell Research.

Kuhn, R. M., D. Haussler, and W. J. Kent. 2013. "The UCSC genome browser and associated tools." *Brief Bioinform* 14 (2):144-61. doi: 10.1093/bib/bbs038.

Lambert, S. A., A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch. 2018. "The Human Transcription Factors." *Cell* 175 (2):598-599. doi: 10.1016/j.cell.2018.09.045.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J.

Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, and International Human Genome Sequencing Consortium. 2001. "Initial sequencing and analysis of the human genome." *Nature* 409 (6822):860-921. doi: 10.1038/35057062.

Larsson, A, B Reinius, T Jacob, T Dalessandri, G Hendriks, M Kasper, and R Sanderg. 2019. Transcriptional bursts explain autosomal random monoallelic expression and affect allelic imbalance. bioRXiv.

Latchman, D. S. 1997. "Transcription factors: an overview." *Int J Biochem Cell Biol* 29 (12):1305-12.

Law, C. W., Y. Chen, W. Shi, and G. K. Smyth. 2014. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts." *Genome Biol* 15 (2):R29. doi: 10.1186/gb-2014-15-2-r29.

Lee, C., S. Patil, and M. A. Sartor. 2016. "RNA-Enrich: a cut-off free functional enrichment testing method for RNA-seq with improved detection power." *Bioinformatics* 32 (7):1100-2. doi: 10.1093/bioinformatics/btv694.

Lee, Christopher T., Cavalcante, G. Raymond, Lee, Qin, Chee, Tingting, Patil, Snehal, Wang, Shuze, Tsai, TY Zing, Boyle, P Alan, Sartor, and A Maureen. 2018. "Poly-Enrich: Count-based Methods for Gene Set Enrichment Testing with Genomic Regions and Updates to ChIP-Enrich."488734.

Li, G., Y. Chen, M. P. Snyder, and M. Q. Zhang. 2017. "ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis." *Nucleic Acids Res* 45 (1):e4. doi: 10.1093/nar/gkw809.

Li, S., C. Wan, R. Zheng, J. Fan, X. Dong, C. A. Meyer, and X. S. Liu. 2019. "Cistrome-GO: a web server for functional enrichment analysis of transcription factor ChIP-seq peaks." *Nucleic Acids Res* 47 (W1):W206-W211. doi: 10.1093/nar/gkz332.

Li, W., D. Notani, and M. G. Rosenfeld. 2016. "Enhancers as non-coding RNA transcription units: recent insights and future perspectives." *Nat Rev Genet* 17 (4):207-23. doi: 10.1038/nrg.2016.4.

Liberzon, A., C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov, and P. Tamayo. 2015. "The Molecular Signatures Database (MSigDB) hallmark gene set collection." *Cell Syst* 1 (6):417-425. doi: 10.1016/j.cels.2015.12.004.

Liberzon, A., A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov. 2011. "Molecular signatures database (MSigDB) 3.0." *Bioinformatics* 27 (12):1739-40. doi: 10.1093/bioinformatics/btr260.

Liu, X., B. Wu, J. Szary, E. M. Kofoed, and F. Schaufele. 2007. "Functional sequestration of transcription factor activity by repetitive DNA." *J Biol Chem* 282 (29):20868-76. doi: 10.1074/jbc.M702547200.

Lizio, M., J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, S. Fukuda, F. Hori, S. Ishikawa-Kato, C. J. Mungall, E. Arner, J. K. Baillie, N. Bertin, H. Bono, M. de Hoon, A. D. Diehl, E. Dimont, T. C. Freeman, K. Fujieda, W. Hide, R. Kaliyaperumal, T. Katayama, T. Lassmann, T. F. Meehan, K. Nishikata, H. Ono, M. Rehli, A. Sandelin, E. A. Schultes, P. A. 't Hoen, Z. Tatum, M. Thompson, T. Toyoda, D. W. Wright, C. O. Daub, M. Itoh, P. Carninci, Y. Hayashizaki, A. R. Forrest, H. Kawaji, and FANTOM consortium. 2015. "Gateways to the FANTOM5 promoter level mammalian expression atlas." *Genome Biol* 16:22. doi: 10.1186/s13059-014-0560-6.

Love, M. I., W. Huber, and S. Anders. 2014. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biol* 15 (12):550. doi: 10.1186/s13059-014-0550-8.

McGinnis, C. S., L. M. Murrow, and Z. J. Gartner. 2019. "DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors." *Cell Syst* 8 (4):329-337.e4. doi: 10.1016/j.cels.2019.03.003.

McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. "UMAP: Uniform Manifold Approximation and Projection." doi: https://doi.org/10.21105/joss.00861.

McLean, C. Y., D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano. 2010. "GREAT improves functional interpretation of cis-regulatory regions." *Nat Biotechnol* 28 (5):495-501. doi: 10.1038/nbt.1630.

Melamed, P., Y. Yosefzon, S. Rudnizky, and L. Pnueli. 2016. "Transcriptional enhancers: Transcription, function and flexibility." *Transcription* 7 (1):26-31. doi: 10.1080/21541264.2015.1128517.

Nguyen, T. A., R. D. Jones, A. R. Snavely, A. R. Pfenning, R. Kirchner, M. Hemberg, and J. M. Gray. 2016. "High-throughput functional comparison of promoter and enhancer activities." *Genome Res* 26 (8):1023-33. doi: 10.1101/gr.204834.116.

Nord, A. S., M. J. Blow, C. Attanasio, J. A. Akiyama, A. Holt, R. Hosseini, S. Phouanenavong, I. Plajzer-Frick, M. Shoukry, V. Afzal, J. L. Rubenstein, E. M. Rubin, L. A. Pennacchio, and A. Visel. 2013. "Rapid and pervasive changes in genome-wide enhancer usage during mammalian development." *Cell* 155 (7):1521-31. doi: 10.1016/j.cell.2013.11.033.

Pennacchio, L. A., W. Bickmore, A. Dean, M. A. Nobrega, and G. Bejerano. 2013. "Enhancers: five essential questions." *Nat Rev Genet* 14 (4):288-95. doi: 10.1038/nrg3458.

Pinsach-Abuin, Mel·lina and Mates Jesus and del Olmo Bernat and Allegue Catarina and Brugada Ramon and Garcia-Bassets Ivan and Pagans Sara. 2016. "Regulome-Seq: Searching for Single Nucleotide Variants (SNVs) Associated with Disease Beyond Protein-Coding Regions." *The FASEB Journal* 30 (1_supplement):1180.4-1180.4. doi: 10.1096/fasebj.30.1_supplement.1180.4.

Pristerà, A., W. Lin, A. K. Kaufmann, K. R. Brimblecombe, S. Threlfell, P. D. Dodson, P. J. Magill, C. Fernandes, S. J. Cragg, and S. L. Ang. 2015. "Transcription factors FOXA1 and FOXA2 maintain dopaminergic neuronal properties and control feeding behavior in adult mice." *Proc Natl Acad Sci U S A* 112 (35):E4929-38. doi: 10.1073/pnas.1503911112.

Qian, J., N. Esumi, Y. Chen, Q. Wang, I. Chowers, and D. J. Zack. 2005. "Identification of regulatory targets of tissue-specific transcription factors: application to retina-specific gene regulation." *Nucleic Acids Res* 33 (11):3479-91. doi: 10.1093/nar/gki658.

Qin, Tingting, Christopher T Lee, Raymond Cavalcante, Peter Orchard, Heming Yao, Hanrui Zhang, Shuze Wang, Snehal Patil, Alan P Boyle, and Maureen A Sartor. A consensus set of enhancer-target gene assignments improves the interpretation of genome-wide regulatory data. In preparation.

Qu, H., and X. Fang. 2013. "A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project." *Genomics Proteomics Bioinformatics* 11 (3):135-41. doi: 10.1016/j.gpb.2013.05.001.

Rao, S. S., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. 2014. "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." *Cell* 159 (7):1665-80. doi: 10.1016/j.cell.2014.11.021.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics* 26 (1):139-40. doi: 10.1093/bioinformatics/btp616.

Robinson, M. D., and G. K. Smyth. 2008. "Small-sample estimation of negative binomial dispersion, with applications to SAGE data." *Biostatistics* 9 (2):321-32. doi: 10.1093/biostatistics/kxm030.

Roider, H. G., T. Manke, S. O'Keeffe, M. Vingron, and S. A. Haas. 2009. "PASTAA: identifying transcription factors associated with sets of co-regulated genes." *Bioinformatics* 25 (4):435-42. doi: 10.1093/bioinformatics/btn627.

Rosenbloom, K. R., T. R. Dreszer, M. Pheasant, G. P. Barber, L. R. Meyer, A. Pohl, B. J. Raney, T. Wang, A. S. Hinrichs, A. S. Zweig, P. A. Fujita, K. Learned, B. Rhead, K. E. Smith, R. M. Kuhn, D. Karolchik, D. Haussler, and W. J. Kent. 2010. "ENCODE whole-genome data in the UCSC Genome Browser." *Nucleic Acids Res* 38 (Database issue):D620-5. doi: 10.1093/nar/gkp961.

Roundtree, I. A., M. E. Evans, T. Pan, and C. He. 2017. "Dynamic RNA Modifications in Gene Expression Regulation." *Cell* 169 (7):1187-1200. doi: 10.1016/j.cell.2017.05.045.

Roy-Engel, A. M., M. L. Carroll, E. Vogel, R. K. Garber, S. V. Nguyen, A. H. Salem, M. A. Batzer, and P. L. Deininger. 2001. "Alu insertion polymorphisms for the study of human genomic diversity." *Genetics* 159 (1):279-90.

Rubtsov, Mikhail A., Yury S. Polikanov, Vladimir A Bondarenko, Yuh-Hwa Wang, and Vasily M. Studitsky. 2006. Chromatin structure can strongly facilitate enhancer action over a distance. Proceedings of the National Academy of Sciences.

Sanyal, A., B. R. Lajoie, G. Jain, and J. Dekker. 2012. "The long-range interaction landscape of gene promoters." *Nature* 489 (7414):109-13. doi: 10.1038/nature11279.

Sasaki, Yutaka. 2007. "The truth of the F-measure." *Teach Tutor Mater*.

Satija, R., J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev. 2015. "Spatial reconstruction of single-cell gene expression data." *Nat Biotechnol* 33 (5):495-502. doi: 10.1038/nbt.3192.

Sato, S., M. Nakamura, D. H. Cho, S. J. Tapscott, H. Ozaki, and K. Kawakami. 2002. "Identification of transcriptional targets for Six5: implication for the pathogenesis of myotonic dystrophy type 1." *Hum Mol Genet* 11 (9):1045-58. doi: 10.1093/hmg/11.9.1045.

Schmidt, D., M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield, and D. T. Odom. 2009. "ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions." *Methods* 48 (3):240-8. doi: 10.1016/j.ymeth.2009.03.001.

Schoenfelder, S., and P. Fraser. 2019. "Long-range enhancer-promoter contacts in gene expression control." *Nat Rev Genet* 20 (8):437-455. doi: 10.1038/s41576-019-0128-0.

Schoenherr, C. J., and D. J. Anderson. 1995. "The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes." *Science* 267 (5202):1360-3. doi: 10.1126/science.7871435.

Segrè, A. V., L. Groop, V. K. Mootha, M. J. Daly, D. Altshuler, DIAGRAM Consortium, and MAGIC investigators. 2010. "Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits." *PLoS Genet* 6 (8). doi: 10.1371/journal.pgen.1001058.

Seth, K. A., and J. A. Majzoub. 2001. "Repressor element silencing transcription factor/neuron-restrictive silencing factor (REST/NRSF) can act as an enhancer as well as a repressor of corticotropin-releasing hormone gene transcription." *J Biol Chem* 276 (17):13917-23. doi: 10.1074/jbc.M007745200.

Sharan, Roded. Lecture 11, January 4, 2007. Analysis of Biological Networks: Transcriptional Networks - Promoter Sequence Analysis. Tel Aviv University . .

Shlyueva, D., G. Stampfel, and A. Stark. 2014. "Transcriptional enhancers: from properties to genome-wide predictions." *Nat Rev Genet* 15 (4):272-86. doi: 10.1038/nrg3682.

Sloan, C. A., E. T. Chan, J. M. Davidson, V. S. Malladi, J. S. Strattan, B. C. Hitz, I. Gabdank, A. K. Narayanan, M. Ho, B. T. Lee, L. D. Rowe, T. R. Dreszer, G. Roe, N. R. Podduturi, F. Tanaka, E. L. Hong, and J. M. Cherry. 2016. "ENCODE data at the ENCODE portal." *Nucleic Acids Res* 44 (D1):D726-32. doi: 10.1093/nar/gkv1160.

Slonim, D. K., and I. Yanai. 2009. "Getting started in gene expression microarray analysis." *PLoS Comput Biol* 5 (10):e1000543. doi: 10.1371/journal.pcbi.1000543.
Solyom, S., and H. H. Kazazian. 2012. "Mobile elements in the human genome: implications for disease." *Genome Med* 4 (2):12. doi: 10.1186/gm311.
Sonenberg, N., and A. G. Hinnebusch. 2009. "Regulation of translation initiation in eukaryotes: mechanisms and biological targets." *Cell* 136 (4):731-45. doi: 10.1016/j.cell.2009.01.042.
Soneson, C., and M. D. Robinson. 2018. "Bias, robustness and scalability in single-cell differential expression analysis." *Nat Methods* 15 (4):255-261. doi: 10.1038/nmeth.4612.
Spitz, F., J. Demignon, A. Porteu, A. Kahn, J. P. Concordet, D. Daegelen, and P. Maire. 1998. "Expression of myogenin during embryogenesis is controlled by Six/sine oculis homeoproteins through a conserved MEF3 binding site." *Proc Natl Acad Sci U S A* 95 (24):14220-5. doi: 10.1073/pnas.95.24.14220.
Stadhouders, R., A. van den Heuvel, P. Kolovos, R. Jorna, K. Leslie, F. Grosveld, and E. Soler. 2012. "Transcription regulation by distal enhancers: who's in the loop?" *Transcription* 3 (4):181-6. doi: 10.4161/trns.20720.
Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. 2005. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." *Proc Natl Acad Sci U S A* 102 (43):15545-50. doi: 10.1073/pnas.0506580102.
Sun, S., J. Zhu, S. Mozaffari, C. Ober, M. Chen, and X. Zhou. 2019. "Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies." *Bioinformatics* 35 (3):487-496. doi: 10.1093/bioinformatics/bty644.
Svensson, V. 2020. "Droplet scRNA-seq is not zero-inflated." *Nat Biotechnol*. doi: 10.1038/s41587-019-0379-5.
Svensson, V., S. A. Teichmann, and O. Stegle. 2018. "SpatialDE: identification of spatially variable genes." *Nat Methods* 15 (5):343-346. doi: 10.1038/nmeth.4636.
Tang, Z., O. J. Luo, X. Li, M. Zheng, J. J. Zhu, P. Szalaj, P. Trzaskoma, A. Magalska, J. Wlodarczyk, B. Ruszczycki, P. Michalski, E. Piecuch, P. Wang, D. Wang, S. Z. Tian, M. Penrad-Mobayed, L. M. Sachs, X. Ruan, C. L. Wei, E. T. Liu, G. M. Wilczynski, D. Plewczynski, G. Li, and Y. Ruan. 2015. "CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription." *Cell* 163 (7):1611-27. doi: 10.1016/j.cell.2015.11.024.
Tarailo-Graovac, M., and N. Chen. 2009. "Using RepeatMasker to identify repetitive elements in genomic sequences." *Curr Protoc Bioinformatics* Chapter 4:Unit 4.10. doi: 10.1002/0471250953.bi0410s25.
Thomas, C. A., A. C. Paquola, and A. R. Muotri. 2012. "LINE-1 retrotransposition in the nervous system." *Annu Rev Cell Dev Biol* 28:555-73. doi: 10.1146/annurev-cellbio-101011-155822.
Thong, Tasha, and Justin Colacino. "Untitled MyoLum AA vs EA manuscript."

Thong, Tasha, Yutong Wang, Michael D. Brooks, Christopher T. Lee, Clayton Scott, Laura Balzano, Max S. Wicha, and Justin A. Colacino. "Single-cell transcriptions of conditionally reprogrammed human mammary cells reveals an enrichment of hybrid epithelial/mesenchymal cells." *In preparation*.

Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos. 2012. "The accessible chromatin landscape of the human genome." *Nature* 489 (7414):75-82. doi: 10.1038/nature11232.

Trapnell, C., D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. 2014. "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells." *Nat Biotechnol* 32 (4):381-386. doi: 10.1038/nbt.2859.

Van den Berge, K., F. Perraudeau, C. Soneson, M. I. Love, D. Risso, J. P. Vert, M. D. Robinson, S. Dudoit, and L. Clement. 2018. "Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications." *Genome Biol* 19 (1):24. doi: 10.1186/s13059-018-1406-4.

van der Maaten, Laurens, and Geoffrey Hinton. 2008. "Visualizing Data using t-SNE."2579-2605.

van Dijk, D., R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, B. Bierie, L. Mazutis, G. Wolf, S. Krishnaswamy, and D. Pe'er. 2018. "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion." *Cell* 174 (3):716-729.e27. doi: 10.1016/j.cell.2018.05.061.

van Heyningen, V., and W. Bickmore. 2013. "Regulation from a distance: long-range control of gene expression in development and disease." *Philos Trans R Soc Lond B Biol Sci* 368 (1620):20120372. doi: 10.1098/rstb.2012.0372.

Vieth, B., S. Parekh, C. Ziegenhain, and et al. 2019. A systematic evaluation of single cell RNA-seq analysis pipelines. Nat Column.

Wang, B., J. M. Cunningham, and X. H. Yang. 2015. "Seq2pathway: an R/Bioconductor package for pathway analysis of next-generation sequencing data." *Bioinformatics* 31 (18):3043-5. doi: 10.1093/bioinformatics/btv289.

Wanichnopparat, W., K. Suwanwongse, P. Pin-On, C. Aporntewan, and A. Mutirangura. 2013. "Genes associated with the cis-regulatory functions of intragenic LINE-1 elements." *BMC Genomics* 14:205. doi: 10.1186/1471-2164-14-205.

Welch, R. P., C. Lee, P. M. Imbriano, S. Patil, T. E. Weymouth, R. A. Smith, L. J. Scott, and M. A. Sartor. 2014. "ChIP-Enrich: gene set enrichment testing for ChIP-seq data." *Nucleic Acids Res* 42 (13):e105. doi: 10.1093/nar/gku463.

Weng, L., F. Macciardi, A. Subramanian, G. Guffanti, S. G. Potkin, Z. Yu, and X. Xie. 2011. "SNP-based pathway enrichment analysis for genome-wide association studies." *BMC Bioinformatics* 12:99. doi: 10.1186/1471-2105-12-99.

Wirth, T., and S. Ylä-Herttuala. 2014. "Gene Therapy Used in Cancer Treatment." *Biomedicines* 2 (2):149-162. doi: 10.3390/biomedicines2020149.

Wittkopp, P. J., and G. Kalay. 2011. "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence." *Nat Rev Genet* 13 (1):59-69. doi: 10.1038/nrg3095.

Wood, Simon N, Yannig Goude, and Simon Shaw. 2015. "Generalized additive models for large data sets." *Journal of the Royal Statistical Society* 64 (1):139-155.

Woodbury, Max A. 1950. *Inverting modified matrices*. Princeton, NJ.

Xie, C., and M. T. Tammi. 2009. "CNV-seq, a new method to detect copy number variation using high-throughput sequencing." *BMC Bioinformatics* 10:80. doi: 10.1186/1471-2105-10-80.

Yoon, S., H. C. T. Nguyen, Y. J. Yoo, J. Kim, B. Baik, S. Kim, and D. Nam. 2018. "Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2." *Nucleic Acids Res* 46 (10):e60. doi: 10.1093/nar/gky175.

Yu, G., L. G. Wang, and Q. Y. He. 2015. "ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization." *Bioinformatics* 31 (14):2382-3. doi: 10.1093/bioinformatics/btv145.

Zhang, K., S. Cui, S. Chang, L. Zhang, and J. Wang. 2010. "i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study." *Nucleic Acids Res* 38 (Web Server issue):W90-5. doi: 10.1093/nar/gkq324.

Zhang, Shulin, Ostap Okhrin, Qian M. Zhou, and Peter X. Song. 2016. "Goodness-of-fit test for specification of semiparametric copula dependence models." *Journal of Econometrics* 193 (1):215-233.